

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



Avanzando en la Modelización de la Exposición al Tratamiento
Farmacológico a través de Algoritmos Computacionales en
Registros Electrónicos de Salud

Doctorando: Dan Ouchi Vernet

Directores: Ferran Torres Benítez y Rosa Morros Pedrós

Tutor académico: Ignasi J. Gich Saladich

Programa de Doctorado en Farmacología

Departamento de Farmacología, de Terapéutica i de Toxicologia

Universitat Autònoma de Barcelona

Barcelona, diciembre 2023



El doctorando Sr. Dan Ouchi Vernet, con DNI: 53288141Q

DECLARA QUE

Es autor de la tesis doctoral titulada “Avanzando en la Modelización de la Exposición al Tratamiento Farmacológico a través de Algoritmos Computacionales en Registros Electrónicos de Salud”.



El Dr. Ferran Torres Benítez, Profesor Agregado e Investigador de la Unidad de Bioestadística, Facultad de Medicina de la Universitat Autònoma de Barcelona, con DNI: 43401371A, la Dra. Rosa Morros Pedrós, Profesora Asociada del Departamento de Farmacologia, de Terapèutica i de Toxicologia, con DNI: 40882722S, como directores, y el Dr. Ignasi J. Gich Saladich, con DNI: 33871528A, como tutor

DECLARAN QUE

La tesis doctoral, con título “Avanzando en la Modelización de la Exposición al Tratamiento Farmacológico a través de Algoritmos Computacionales en Registros Electrónicos de Salud”, presentada por Dan Ouchi Vernet para obtener el título de Doctor, ha sido realizada bajo nuestra supervisión.

Abstract

This thesis consists of a compilation of three research articles produced between 2020 and 2023. The three articles share a common thread: they are distinct contributions based on the method developed during the thesis to improve and automate the modeling of drug exposure using electronic health records (EHR). In essence, the algorithm developed, called “smooth”, is inspired by statistical time series techniques that calculate central trends using moving averages. This technique is mainly used to obtain a more stable numerical series, free of trends and seasonalities, and to “smooth” out short-term fluctuations. Using this strategy, a specific method was created for EHR, which automatically calculates the most likely exposure profile over time.

In the first work of this thesis, the creation, development, implementation and subsequent validation of the algorithm is presented in detail. Beyond the details of its operation, one of the features of the algorithm is that it incorporates a tuning parameter called Wt , which controls the level of smoothing. We evaluated the influence of this parameter through a simulation study. As we increased this parameter, the original data were modified, although the algorithm affects few patients, and most of them show changes of less than 5%. After developing and implementing the algorithm, we sought to analyze its usefulness from a clinical and pharmacological perspective. For this purpose, we conducted a validation with the participation of four independent experts, who evaluated treatment profiles in different pharmacological scenarios. Each reviewer, in a blinded manner, compared the results obtained by the algorithm with respect to the original data, as well as in relation to a non-automated method more common in the literature. The validation showed a positive evaluation in favor of the “smooth” algorithm in overall and across the different scenarios. In addition, significant differences were found in comparison with the traditional method.

In the second article, we explored the main advantages of using the algorithm to analyze longitudinal treatment patterns in newly diagnosed patients with type 2 diabetes mellitus (T2DM). Our main objective was to study how treatments prescribed in primary care do or do not adhere to clinical guidelines. To achieve this, we included all cases registered in primary care settings between 2015 and 2020 in Catalonia, a total of 86,854 patients. Using the algorithm, we modeled longitudinal patterns up to third-line treatments, identifying possible changes, additions, deletions, and discontinuations throughout the study period. In addition, the method was useful for calculating exposure times, compliance and correlating treatment patterns with glycosylated hemoglobin

(HbA1c) values. The main results indicated that 78.3% initiated antidiabetic treatment with single therapy, with metformin being the predominant therapeutic option in first and third line, while combination therapy (CT) of metformin with inhibitors of dipeptidyl peptidase 4 (DPP4i) or metformin with sulfonylurea was more prevalent in second line. The most common longitudinal pattern consisted of starting metformin for 15 months, introducing a second antidiabetic agent for 6 months in second-line treatment, and finally returning to metformin treatment. Treatment patterns were significantly associated with HbA1c values. High levels ($\geq 8\%$) favored switches to CT, whereas low levels led to switches to monotherapy or temporary discontinuation of treatment.

In the third article of the thesis, we address another study in patients diagnosed with T2DM, with the aim of analyzing the impact of adding a second-line treatment to patients under treatment with metformin who fail to control the disease. We analyzed a total of 28,425 patients between 2015 and 2020, using the algorithm to determine the timing of addition of second-line treatment, calculate exposure time and assess compliance. The study variable was time to glycosylated hemoglobin control ($\text{HbA1c} < 8\%$) expressed as restricted mean survival time (RMST), and differences in RMST of each treatment relative to the combination of metformin with sulfonylurea were compared. In summary, all treatments significantly reduced HbA1c, although combinations of metformin with glucagon-like peptide-1 (GLP1) or metformin with insulin reached the target earlier ($P < .001$). We observed good compliance in all study groups, except for the combination with insulin, and adherence resulted in a significant reduction in time to HbA1c control, with the exception of patients under the combination of metformin with sodium-glucose transport protein 2 inhibitors (SGLT2i).

To obtain the results presented in this thesis, we used the Sistema de Información para la Investigación en Atención Primaria (SIDIAP) as our main data source.

Resum

Aquesta tesi consisteix en una recopilació de tres articles d'investigació elaborats entre el 2020 i el 2023. Els tres articles comparteixen un lligam comú: són diferents contribucions basades en el mètode desenvolupat durant la tesi per millorar i automatitzar la modelització de l'exposició a fàrmacs utilitzant la història clínica informatitzada (EHR). En essència, l'algoritme desenvolupat, anomenat “smooth”, s'inspira en tècniques estadístiques de sèries temporals que calculen tendències centrals mitjançant mitjanes mòbils. Aquesta tècnica s'emplea principalment per obtenir una sèrie numèrica més estable, lliure de tendències i estacionalitats, i per suavitzar les fluctuacions de curta durada. Utilitzant aquesta estratègia, es va crear un mètode específic per als EHR, el qual calcula automàticament el perfil d'exposició més probable al llarg del temps.

En el primer treball d'aquesta tesi, es presenta en detall la creació, desenvolupament, implementació i posterior validació de l'algorisme. Més enllà dels detalls del seu funcionament, una de les característiques de l'algorisme és que incorpora un paràmetre d'ajust anomenat Wt , que controla el grau d'alleujament. Vam avaluar l'impacte d'aquest paràmetre mitjançant un estudi de simulació. A mesura que augmentem aquest paràmetre, les dades originals es modifiquen, encara que l'algorisme afecta en pocs pacients, i la majoria d'ells presenta canvis inferiors al 5%. Després de desenvolupar i implementar l'algorisme, vam buscar analitzar la seva utilitat des d'una perspectiva clínica i farmacològica. Amb aquest propòsit, vam dur a terme una validació amb la participació de quatre experts independents, els quals van avaluar perfils de tractament en diferents escenaris farmacològics. Cada revisor, de manera cega, va comparar els resultats obtinguts per l'algorisme respecte a les dades originals, així com en relació a un mètode no automàtic més comú a la literatura. La validació va evidenciar una avaluació positiva a favor de l'algorisme “smooth” en general i per als diferents escenaris. Addicionalment, es van trobar diferències significatives en comparació amb el mètode tradicional.

En el segon article, vam aprofitar els principals avantatges d'utilitzar l'algorisme per analitzar els patrons longitudinals de tractament en pacients diagnosticats recentment amb diabetis mellitus tipus 2 (T2DM). El nostre objectiu principal va ser estudiar com els tractaments prescrits en atenció primària s'adhereixen o no a les guies clíniques. Per aconseguir-ho, vam incloure tots els casos registrats en entorns d'atenció primària entre el 2015 i el 2020 a Catalunya, un total de

86.854 pacients. Mitjançant l'algorisme, vam modelar els patrons longitudinals fins als tractaments de tercera línia, identificant possibles canvis, addicions, eliminacions i interrupcions al llarg del període d'estudi. A més, el mètode va ser útil per calcular temps d'exposició, el compliment i correlacionar patrons de tractament amb els valors de hemoglobina glicada (HbA1c). Els resultats principals van indicar que el 78,3% va iniciar el tractament antidiabètic amb teràpia simple, sent la metformina l'opció terapèutica predominant en primera i tercera línia, mentre que la teràpia combinada (CT) de metformina amb DPP4i o metformina amb sulfonilurea va ser més prevalent en segona línia. El patró longitudinal més comú va consistir en iniciar metformina durant 15 mesos, introduir un segon agent antidiabètic durant 6 mesos en el tractament de segona línia i, finalment, tornar al tractament amb metformina. Els patrons de tractament es van associar significativament als valors de HbA1c. Nivells alts ($\geq 8\%$) van afavorir canvis a CT, mentre que nivells baixos van dur a canvis a monoteràpia o interrupció temporal del tractament.

En el tercer article de la tesi, abordem un altre estudi en pacients diagnosticats amb T2DM, amb l'objectiu d'analitzar l'impacte d'afegir un tractament de segona línia a pacients sota tractament amb metformina que no aconsegueixen controlar la malaltia. Vam analitzar un total de 28.425 pacients entre el 2015 i el 2020, utilitzant l'algorisme per determinar el moment d'addició del tractament de segona línia, calcular el temps d'exposició i avaluar el compliment. La variable d'estudi va ser el temps fins al control de la hemoglobina glicada ($\text{HbA1c} < 8\%$) expressat com a temps de supervivència mig restringit (RMST), i es van comparar les diferències en el RMST de cada tractament en relació amb la combinació de metformina amb sulfonilurea. En resum, tots els tractaments van reduir significativament la HbA1c, tot i que les combinacions de metformina amb un pèptid semblant al glucagó-1 (GLP1) o metformina amb insulina van assolir l'objectiu abans ($P < 0,001$). Vam observar un bon compliment en tots els grups d'estudi, excepte en la combinació amb insulina, i l'adherència va resultar en una reducció significativa del temps fins al control de la HbA1c, llevat dels pacients sota la combinació de metformina amb inhibidors del cotransportador de sodi-glucosa tipus 2 (SGLT2i).

Per obtenir tots els resultats presentats en aquesta tesi, vam fer ús del Sistema d'Informació per a la Investigació en Atenció Primària (SIDIAP) com a principal font de dades de referència.

Resumen

Esta tesis consiste en una recopilación de tres artículos de investigación elaborados entre 2020 y 2023. Los tres artículos comparten un vínculo común: son distintas contribuciones basadas en el método desarrollado durante la tesis para mejorar y automatizar la modelización de la exposición a medicamentos utilizando la historia clínica electrónica (EHR, por sus siglas en inglés). En esencia, el algoritmo desarrollado, denominado “smooth”, se inspira en técnicas estadísticas de series temporales que calculan tendencias centrales mediante medias móviles. Esta técnica se emplea principalmente para obtener una serie numérica más estable, libre de tendencias y estacionalidades, y para suavizar las fluctuaciones de corta duración. Utilizando esta estrategia, se creó un método específico para los EHR, el cual calcula automáticamente el perfil de exposición más probable a lo largo del tiempo.

En el primer trabajo de esta tesis, se presenta en detalle la creación, desarrollo, implementación y posterior validación del algoritmo. Más allá de los detalles de su funcionamiento, una de las características del algoritmo es que incorpora un parámetro de ajuste llamado W_t , que controla el grado de suavizado. Evaluamos el impacto de este parámetro mediante un estudio de simulación. A medida que aumentamos dicho parámetro, los datos originales se modifican, aunque el algoritmo afecta en pocos pacientes, y la mayoría de ellos presenta cambios inferiores al 5%. Después de desarrollar y implementar el algoritmo, buscamos analizar su utilidad desde una perspectiva clínica y farmacológica. Con este propósito, llevamos a cabo una validación con la participación de cuatro expertas independientes, quienes evaluaron perfiles de tratamiento en distintos escenarios farmacológicos. Cada revisora, de forma ciega, comparó los resultados obtenidos por el algoritmo con respecto a los datos originales, así como en relación con un método no automático (tradicional) común en la literatura. La validación evidenció una evaluación positiva en favor del algoritmo “smooth” en general y para los diferentes escenarios. Adicionalmente, se encontraron diferencias significativas en comparación con el método tradicional.

En el segundo artículo, aprovechamos las ventajas principales de utilizar el algoritmo para analizar los patrones longitudinales de tratamiento en pacientes recién diagnosticados con diabetes mellitus tipo 2 (T2DM). Nuestro objetivo principal fue estudiar cómo los tratamientos prescritos en atención primaria se adhieren o no a las guías clínicas. Para lograrlo, incluimos todos los casos

registrados en entornos de atención primaria entre 2015 y 2020 en Cataluña, un total de 86,854 pacientes. Mediante el algoritmo, modelamos los patrones longitudinales hasta los tratamientos de tercera línea, identificando posibles cambios, adiciones, deleciones, interrupciones y finalizaciones a lo largo de todo el período de estudio. Además, el algoritmo resultó de utilidad para calcular tiempos de exposición, el cumplimiento y correlacionar patrones de tratamiento con los valores de hemoglobina glicada (HbA1c). Los resultados principales indicaron que el 78.3% inició el tratamiento antidiabético con terapia simple, siendo la metformina la opción terapéutica predominante en primera y tercera línea, mientras que la terapia combinada (CT) de metformina con DPP4i o metformina con sulfonilurea fue más prevalente en segunda línea. El patrón longitudinal más común consistió en iniciar metformina durante 15 meses, introducir un segundo agente antidiabético durante 6 meses en el tratamiento de segunda línea y, finalmente, volver al tratamiento con metformina. Los patrones de tratamiento se asociaron significativamente a los valores de HbA1c. Niveles altos ($\geq 8\%$) favorecieron cambios a CT, mientras que niveles bajos llevaron a cambios a monoterapia o interrupción temporal del tratamiento.

En el tercer artículo de la tesis, abordamos otro estudio en pacientes diagnosticados con T2DM, con el objetivo de analizar el impacto de añadir un tratamiento de segunda línea pacientes bajo tratamiento con metformina que no logran controlar la enfermedad. Analizamos un total de 28,425 pacientes entre 2015 y 2020, utilizando el algoritmo para determinar el momento de adición del tratamiento de segunda línea, calcular el tiempo de exposición y calcular el cumplimiento. La variable de estudio fue el tiempo hasta el control de la hemoglobina glicada ($\text{HbA1c} < 8\%$) expresado como tiempo de supervivencia medio restringido (RMST), y se compararon las diferencias en RMST de cada tratamiento en relación a la combinación de metformina con sulfonilurea. En resumen, todos los tratamientos redujeron significativamente la HbA1c, aunque las combinaciones de metformina con péptido similar al glucagón-1 (GLP1) y de metformina con insulina alcanzaron el objetivo antes. Observamos un buen cumplimiento en todos los grupos de estudio, excepto en la combinación con insulina, y la adherencia resultó en una reducción significativa del tiempo hasta el control de la HbA1c, con excepción de los pacientes bajo la combinación de metformina con inhibidores del cotransportador de sodio-glucosa tipo 2 (SGLT2i).

Para obtener los resultados presentados en esta tesis, hicimos uso del Sistema de Información para la Investigación en Atención Primaria (SIDIAP) como nuestra base de datos de EHR de referencia.

Agradecimientos

Este proyecto inició hace varios años, y a lo largo de su desarrollo, he compartido experiencias con diversas personas a las que quiero expresar mi agradecimiento por su apoyo y contribuciones, ya sea de forma directa o indirecta, a esta tesis.

En primer lugar, me gustaría expresar mi más sincero agradecimiento al Dr. Ferran Torres y a la Dra. Rosa Morros, quienes han dirigido esta tesis. Ferran, gracias por tus continuas sugerencias y por dedicarme tu tiempo a pesar de tus múltiples responsabilidades; tus conocimientos en estadística y tu perspectiva fueron invaluable y fundamentales para el éxito del avance de la tesis. Rosa, agradezco tu introducción a este vasto e intrigante campo de la farmacoepidemiología en bases de datos electrónicos. He valorado enormemente todos tus buenos consejos, tu paciencia y tu conocimiento en el ámbito de los medicamentos. He aprendido mucho de ti y disfruté participando en discusiones y generando ideas para futuras líneas de investigación. Me siento afortunado de haberlos tenido a ambos como directores, tanto en el ámbito académico como personal.

Un agradecimiento especial a todo el grupo de investigación de la UEM y a todo/as las investigador/as y coautor/as de los artículos. María, Ainhoa, Cristina, Carles, Ramón y Ana por todo la ayuda, colaboración, consejos y tiempo que le habéis dedicado, enriqueciendo significativamente el trabajo publicado. También al resto del equipo por abrirme las puertas al grupo de investigación, por todo el apoyo profesional y personal, y los buenos momentos que hemos vivido estos años. He aprendido mucho de cada uno de vosotros/as y espero que sigamos investigando juntos en el futuro.

Mi sincero agradecimiento también a Iván Galván y Álvaro Franquet por la oportunidad de aprender con vosotros y trabajar juntos en la tesis y en la programación del paquete *smoothy*. Gracias al grupo de 'Analytics' por toda la excelente carrera laboral que estamos viviendo y por todos los buenos momentos disfrutados. Quiero expresar mi gratitud a mis amigos del Máster MEIO y de la carrera, así como a mi 'compañero' Xavi, con quien he compartido incontables momentos memorables hablando de la vida, del deporte que amamos y, por supuesto, de estadística. Un especial agradecimiento a mis hermanos Urban, que siempre están y estarán ahí, enfrentando lo que la vida nos presente.

Finalmente, quiero dedicar un agradecimiento especial a mi compañera Jessica por dar sentido al día a día y por el amor y apoyo recibidos a lo largo de estos años. A mis padres, tanto los de aquí como los de allá, les agradezco su comprensión y aliento en esta aventura. A todas y a todos, gracias.

Motivación

El trabajo de investigación desarrollado durante esta tesis ha sido promovido por la 'Unitat de Recerca del Medicament' (UEM¹), un grupo de investigación que forma parte del Instituto de Investigación en Atención Primaria de Salud Jordi Gol (IDIAPJGol²). La UEM tiene como responsabilidad el diseño, coordinación y supervisión de la investigación en medicamentos, lo cual incluye brindar apoyo a estudios observacionales de farmacoepidemiología, farmacovigilancia o estudios de posautorización con medicamentos (PASS) mediante el uso de datos del mundo real (RWD, del inglés *Real-World Data*), principalmente usando el Sistema de Información para la Investigación en Atención Primaria (SIDIAP), la historia clínica electrónica (EHR, por su sigla en inglés *Electronic Health Records*) en atención primaria desarrollada y mantenida en IDIAPJGol. En este contexto, esta tesis se centra en investigar nuevas formas de mejorar el análisis a la exposición a medicamentos utilizando la EHR.

¹ <https://www.idiapjgol.org/index.php/ca/area-cientifica/49-recerca/uem/132-unitat-d-estudi-del-medicament-uem>

² <https://www.idiapjgol.org/index.php/ca/>

Estructura de la Tesis

La memoria de la tesis doctoral está estructurada de la siguiente manera: El capítulo 1 es la introducción. Los capítulos 2 y 3 describen la hipótesis y los objetivos de esta tesis. El capítulo 4 está formado por el compendio de los artículos publicados y presentados. El capítulo 5 sintetiza los principales resultados obtenidos y el capítulo 6 contiene la discusión.

La tesis concluye con las conclusiones y las futuras líneas de investigación.

Índice de Contenido

Abstract.....	vii
Resum	ix
Resumen	xi
Agradecimientos.....	xiii
Motivación.....	xv
Estructura de la Tesis.....	xvi
Índice de Contenido.....	xvii
Lista de Tablas.....	xxi
Lista de Figuras	xxii
Acrónimos	xxiv
1. Introducción.....	1
1.1 Antecedentes.....	1
1.2 Bases de Datos de EHR.....	10
Aprobación Ética	11
1.3 Datos de Medicamentos	11
Prescripciones Electrónicas	12
Facturaciones en Farmacias.....	12
Visualización de los Datos de Medicamentos.....	12
1.4 Analizar la Exposición a Medicamentos Usando Bases de Datos	14
1.5 Modelar el Tratamiento Longitudinal	14
Inicio del Tratamiento	15
Monoterapia y Terapia Combinada.....	15

Cambios de Tratamiento	15
Adición, Delección y Interrupción Temporal	15
Finalización del Tratamiento.....	16
Duración	16
1.6 Calcular el Cumplimiento.....	17
1.7 Tratamiento de la Diabetes Tipo 2	17
Antidiabéticos.....	17
Patrones Longitudinales de Tratamiento Antidiabético	20
Datos de Hemoglobina Glicada.....	21
Uso en Estudios Observacionales.....	22
1.8 La Influencia de la Calidad de los Datos	23
1.9 Limitaciones de los Datos de Medicamentos en SIDIAP	24
2. Hipótesis	31
3. Objetivos.....	33
4. Compendio de Artículos.....	35
4.1 JMIR Medical Informatics	35
4.2 Diabetes Research and Clinical Practice	47
4.3 Drugs – Real World Outcomes.....	59
5. Resumen y Análisis de Resultados del Tratamiento de la T2DM.....	73
5.1 Descripción Longitudinal de los Tratamientos Antidiabéticos	73
Tratamiento Inicial	73
Monoterapia y Terapia Combinada	75
Patrones de Tratamiento Longitudinal	76
Asociación con la HbA1c.....	78
5.2 Impacto del Tratamiento de Segunda Línea en la HbA1c.....	79
Descripción de la Población	79

Resultados Principales.....	80
6. Discusión	83
6.1 El Algoritmo <i>smooth</i>	83
Algoritmo para Determinar el Tratamiento Más Probable.....	83
Validación del Algoritmo	86
Ventajas del Algoritmo	87
Visualización de los Tratamientos.....	88
Software para el Paquete Estadístico R.....	89
6.2 Modelizar la Exposición a Medicamentos a partir del Algoritmo	89
Patrones Longitudinales de Tratamiento con Antidiabéticos.....	89
Analizar Tratamientos Combinados	89
Selección de Pacientes.....	89
6.3 Limitaciones	90
Limitaciones del Algoritmo.....	90
Limitaciones en los Análisis de Exposición a Fármacos.....	92
Registros de Medicamentos.....	92
Registros de Diagnósticos	92
Limitaciones con el Tiempo de Seguimiento de los Pacientes	93
Limitaciones de los Datos en Pacientes con T2DM.....	93
7. Conclusiones.....	95
8. Nuevas Líneas de Investigación	97
9. Referencias	99
10. Apéndices	111
Apéndice 1. EHR en SIDIAP	111
Tablas y Datos Disponibles en SIDIAP	111
Pacientes con T2DM en SIDIAP.....	115

Apéndice 2. Validación del Algoritmo	121
Validación Cualitativa	121
Concordancia entre Revisoras	125
Apéndice 3. Paquete de R <i>smoothy</i>	127
El Repositorio CRAN.....	127
Contribuir a CRAN:	127
Tutorial del Paquete de R <i>smoothy</i>	128

Lista de Tablas

Tabla 1.	Fuentes Comunes de RWD	2
Tabla 2.	Uso de los EHR en Investigación con Medicamentos.....	4
Tabla 3.	Ejemplo de Registros Electrónicos de Diagnósticos en SIDIAP	10
Tabla 4.	Descripción (en inglés) de Códigos ATC-7 y Grupos de Antidiabéticos (ADM).....	18
Tabla 5.	Otras Dificultades de Analizar las EHR Para Obtener Evidencias Científicas	28
Tabla 6.	Relación entre Objetivos de la Tesis Doctoral y los Trabajos Publicados	33
Tabla 7.	Descripción de los Tratamientos de Primera a Tercera Línea en Función de Edad, HbA1c y Cumplimiento.....	75
Tabla 8.	Características Demográficas y Clínicas de la Población	80
Tabla 9.	Resultados de la Simulación con Diferentes Valores de Wt	86
Tabla Sup. 1.	Descripción de las Tablas y Variables Disponibles en SIDIAP.....	111
Tabla Sup. 2.	Códigos ICD-10 que Identifican a los Pacientes con T2DM	115
Table Sup. 3.	Características de los Pacientes con T2DM en SIDIAP.....	117
Tabla Sup. 4.	Descripción de los Grupos de Medicamentos Analizados en la Validación	122
Tabla Sup. 5.	Ejemplo de Tabla de Validación Completada por Cada Revisora	123
Tabla Sup. 6.	Resultados de la Validación en los Diferentes Grupos de Medicamento.....	124

Lista de Figuras

Figura 1.	Los 10 Países con Mayor Adopción de EHR.....	7
Figura 2.	Población Catalana Representada en SIDIAP	9
Figura 3.	Representación Gráfica de los Datos de Medicamentos en EHR.....	13
Figura 4.	Eventos que Componen la Modelización Longitudinal del Tratamiento.....	16
Figura 5.	Cálculo del Cumplimiento.....	17
Figura 6.	Patrones Longitudinales en ADM	21
Figura 7.	Asociación de HbA1c con los Cambios en la Pauta de Tratamiento	22
Figura 8.	Esquema de Utilización de HbA1c en el Estudio de Pacientes con T2DM	23
Figura 9.	Discrepancias Entre la Realidad y los Registro Electrónicos en SIDIAP	25
Figura 10.	Solapamientos en las Prescripciones Electrónicas	26
Figura 11.	Tratamientos Combinados y Perfiles Irregulares	26
Figura 12.	Formas de Afrontar los Perfiles de Tratamiento Irregulares	27
Figura 13.	Diagrama de Flujo de los Pacientes en Estudio.....	74
Figura 14.	Flujo de Pacientes de Primera a Tercera Línea	76
Figura 15.	Patrones de Tratamiento Longitudinal	78
Figura 16.	Asociación de los Patrones Longitudinales a Cambios en HbA1c.....	79
Figura 17.	Kaplan-Meier y RMST para el Tiempo a HbA1c <8%.....	81
Figura 18.	Medias Móviles en Series Temporales.....	84
Figura 19.	Ejemplos de Usar el “smooth” en Diferentes Medicamentos	85
Figura 20.	Gráfico de Conjuntos (<i>Upset</i>) y Diagrama Aluvial.....	88
Figura 21.	Perfil de un Paciente Analizado en el Tercer Artículo de Investigación.....	90
Figura Sup. 1.	Ejemplo de un Paciente Analizado en el Estudio de Validación	123

Figura Sup. 2.	Concordancia entre las Revisoras Durante la Validación	125
----------------	--	-----

Acrónimos

A

ADM - Medicación Antidiabética, 17

AGR - Agrupador, 10

APAC - Asia-Pacífico, 6

ARA-II - antagonistas de los receptores AT1 de la angiotensina II, 86

C

CDM - Modelo de Datos Comunes, 7

CEIm - Comité Ético de Investigación con Medicamentos, 11

COVID-19 - Enfermedad del Coronavirus, 6

CRAN - Comprehensive R Archive Network, 127

CT - Terapia Combinada, 14

D

DPP4i - Inhibidores de la DPP-4, 18

E

EAP - Equipos de Atención Primaria, 8

eCAP - estación Clínica de Atención Primaria, 8

EHDS - European Health Data Space, 8

EHR - Historia Clínica Electrónica, xv

EMA - Agencia Europea del Medicamento, 7

EPOC - Enfermedad Pulmonar Obstructiva Crónica, 14

G

GDPR - Reglamento General de Protección de Datos, 7

GLP1 - Análogos del GLP-1, 18

I

ICD-10 - Código Diagnóstico ICD-10, 10

ICS - Instituto Catalán de la Salud, 8

IDIAPJGol - Instituto de Investigación en Atención Primaria de Salud Jordi Gol, xv

idp - identificador seudonimizado, 10

IECA - inhibidores de la ECA, 86

L

LIC - Países con Recursos Limitados, 6

M

MA - Medias Móviles, 83

MPR - Tasa de Posesión del Medicamento, 17

N

NDDD - Número de Dosis Diaria Definida, 12

P

PADRIS - Programa de Analítica de Datos para la Investigación y la Innovación en Salud, 8

PASS - Posautorización con Medicamentos, xv

PC - Atención Primaria, 10

SMA - Media Móvil Simple, 83

R

T

RCT - Ensayo Controlado Aleatorizado, 1

T2DM - Diabetes Mellitus Tipo 2, 14

RWD - Datos del Mundo Real, xv

U

RWE - Evidencia el Mundo Real, 1

UE - Unión Europea, 6

S

UEM - Unitat de Recerca del Medicament, xv

SGLT2i - Inhibidores del SGLT2, 18

W

SISAP - Sistema de Información de los Servicios de
Atención Primaria, 8

Wt - Ventana de Suavizado, 85

SISCAT - Sistema Sanitario Público de Cataluña, 8

1. Introducción

1.1 Antecedentes

La evidencia del mundo real (RWE, en inglés *Real-World Evidence*) en medicina se refiere a la evidencia clínica obtenida del análisis de datos reales para comprender el uso, los beneficios y los riesgos de los productos médicos en la práctica clínica real [1]. A diferencia de las pruebas obtenidas a partir de ensayos clínicos, que generalmente involucran a muestras de pacientes seleccionadas bajo condiciones controladas, como el ensayo controlado aleatorizado (RCT, por sus siglas en inglés *Randomized Clinical Trial*), la RWE se obtienen a partir de observaciones y datos recogidos en entornos del mundo real, y suele abarcar una variedad más amplia de pacientes, incluidos aquellos que podrían no haber sido representados en RCT debido a factores como comorbilidades, edad o ubicación geográfica [2]. Este tipo de evidencia permite a los investigadores y profesionales de la salud evaluar el desempeño de los productos médicos y su impacto en los resultados de los pacientes en diversos escenarios.

Estas evidencias pueden generarse a través de diversos diseños de estudio, como estudios observacionales retrospectivos, prospectivos o ensayos pragmáticos, y se derivan del análisis de grandes bases de datos de diferentes fuentes como son la historia clínica electrónica, actividades de reclamaciones y facturación, registros de productos y enfermedades, etc. que conforman lo que denominamos los datos del mundo real (RWD, por sus siglas en inglés *Real World Data*).

Definimos los RWD como información relacionada con la salud del paciente, su experiencia o la prestación de atención médica, recopilada fuera del contexto de un RCT [3]. Los datos del mundo real pueden registrarse de manera rutinaria durante la prestación de atención médica y/o social, o prospectivamente para abordar una o más preguntas de investigación específicas. La mayoría de las fuentes de datos del mundo real son observacionales (o no intervencionales), es decir, no son determinadas por un protocolo de estudio y su origen puede darse en un ámbito local, nacional o internacional [4]. Dichas fuentes se clasifican según la información (cuantitativa o cualitativa) que la conformen (véase Tabla 1) y generalmente incluyen la demografía de los pacientes, el historial médico, el uso de medicamentos, los resultados clínicos (incluidos los resultados informados por los pacientes), las experiencias de los pacientes o usuarios,

el uso de recursos, la genética u ómica, mediciones de laboratorio, el texto libre, los resultados o imágenes de pruebas y los datos autogenerados por los pacientes.

Si bien cada tipo de fuente y de datos tiene sus fortalezas y debilidades, el uso de una u otra dependerá de la pregunta de investigación específica a responder y, a menudo, suelen combinarse para mejorar la calidad y ampliar su cobertura, lo que permite responder a preguntas adicionales.

Tabla 1. Fuentes Comunes de RWD

FUENTE DE DATOS	DESCRIPCIÓN	EJEMPLO
Historia Clínica Electrónica (EHR)	Registros digitales de pacientes almacenadas en soporte electrónico con toda la información referente al paciente y su atención clínica.	Clinical Practice Research Datalink (CPRD) incluye información demográfica y clínica de pacientes inscritos en consultas médicas generales participantes en todo el Reino Unido.
Datos Administrativos	Datos recopilados para fines administrativos por servicios de salud y cuidado social.	El CMBD-HA , contiene datos sobre los procedimientos y diagnósticos realizados en los pacientes ingresados en los hospitales del CatSalut.
Datos de Reclamaciones	Datos sobre el uso de servicios de atención médica, recopilados en sistemas basados en seguros.	Centers for Medicare & Medicaid Services data información sobre personas que reciben servicios de Medicare derivada de información de reembolsos o pago de facturas.
Registros de Pacientes	Sistemas que recopilan datos uniformes para identificar resultados específicos en una población definida.	GCAT , una base de datos diseñada para el estudio de los factores genéticos y ambientales que conducen a la aparición de las enfermedades crónicas en la población general.

Datos Generados por Pacientes	Datos generados directamente por pacientes o sus cuidadores, como dispositivos médicos portátiles y aplicaciones móviles.	Datos autoreportados sobre la infección por COVID-19 o síntomas de long-COVID a partir de una aplicación para teléfonos inteligentes, radar Covid .
Revisiones de Expedientes	Datos extraídos retrospectivamente de la revisión de expedientes de salud de pacientes.	Revisión de historias clínicas en estudios de enfermedades raras para modelar la historia natural de la enfermedad y las vías de tratamiento (Garbade et al, 2021) [5].
Auditoría y Evaluación de Servicios	Auditorías clínicas para comparar la atención actual con mejores prácticas, y evaluaciones de servicios para definir y juzgar la atención actual.	Programas nacionales de auditoría clínica, como el Sentinel Stroke National Audit Programme (SSNAP).
Cohortes Observacionales con Recopilación de Datos Primarios	Estudios prospectivos para responder preguntas de investigación.	EMBRACE-I es un estudio de cohortes prospectivo multicéntrico para evaluar el control local de tumores y la morbilidad para tumores cervicales localmente avanzados.
Encuestas de Salud y Entrevistas	Encuestas, entrevistas y grupos de enfoque para recopilar datos sobre salud y experiencias de pacientes.	La Encuesta Nacional de Salud de España es un informe publicado periódicamente por el Ministerio de Sanidad.

Entre las diversas fuentes de RWD, la Historia Clínica Electrónica destaca como una de las más utilizadas debido a su capacidad para proporcionar datos longitudinales, actualizados y a un bajo coste económico. En términos generales, las EHR se definen como datos longitudinales recopilados de manera electrónica durante la atención médica habitual [6], y surgieron con el propósito de mejorar la calidad de la atención médica y agilizar la gestión de datos de facturación [7].

En la industria farmacéutica, las EHR se utilizan de forma habitual para llevar a cabo investigaciones de mercado, realizar farmacovigilancia, validar marcadores clínicos y evaluar la seguridad de los medicamentos. Sin embargo, debido a que los EHR proporcionan datos observacionales de una extensa población a lo largo de períodos prolongados, su uso en la investigación ha experimentado un crecimiento notable en los últimos años y está teniendo sobre todo un impacto significativo en estudios con medicamentos [8], [9]. Actualmente, se aprovechan para obtener un conocimiento más profundo sobre cómo los medicamentos impactan a los pacientes a través de cambios en el diagnóstico, la progresión de enfermedades y las mediciones de laboratorio. Sin embargo, existe un creciente interés en explorarlos para la búsqueda de nuevos medicamentos, la identificación de relaciones entre enfermedades, el reposicionamiento de medicamentos y el descubrimiento de asociaciones entre genotipo y fenotipo [10] (Tabla 2).

Tabla 2. Uso de los EHR en Investigación con Medicamentos

TIPO DE ESTUDIO	DESCRIPCIÓN	EJEMPLO
Farmacovigilancia	Monitoreo y evaluación del perfil de seguridad de los medicamentos para detectar y gestionar efectos adversos.	Seguimiento de casos de reacciones adversas a un medicamento recién aprobado en una población de pacientes [11].
Eficacia	Comparar la eficacia y seguridad de diferentes opciones de tratamiento para condiciones específicas en entornos del mundo real.	Evaluar la eficacia relativa de dos medicamentos comunes para el control de la hipertensión en pacientes con distintos perfiles de salud [12].
Vigilancia Post-Marketing	Monitorización de los medicamentos después de su aprobación y comercialización al mercado para asegurar su seguridad y eficacia.	Seguimiento de la seguridad y eficacia a largo plazo de un nuevo tratamiento oncológico [13].
Reposicionamiento de Medicamentos	Identificación de nuevos usos para medicamentos existentes y encontrar posibles beneficios	Identificar que un medicamento para la artritis también puede ser

terapéuticos en diferentes condiciones.

eficaz en el tratamiento de la psoriasis [14].

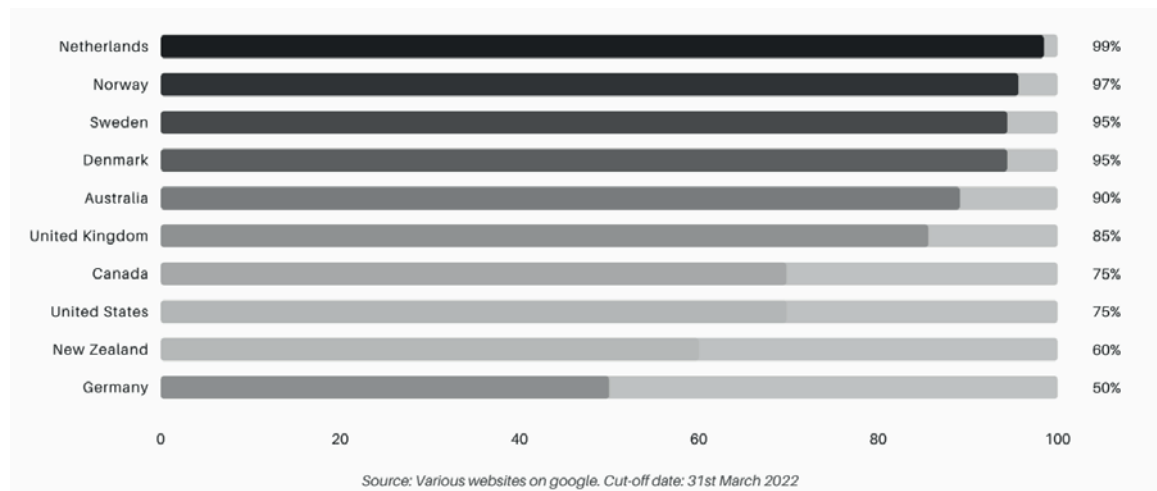
Control de Enfermedades y Detección de Brotes	Utilización de datos en tiempo real de EHR para rastrear la propagación de enfermedades, identificar brotes y evaluar el impacto de intervenciones.	Seguimiento de casos de COVID-19 en tiempo real y detección de áreas con altas tasas de transmisión [15].
Marcadores Clínicos y Biomarcadores	Evaluación de la validez de los marcadores clínicos y biomarcadores para ayudar en el diagnóstico y selección de tratamientos.	Confirmar la utilidad diagnóstica de un biomarcador específico para la detección temprana de cáncer de pulmón [16].
Adherencia y Cumplimiento del Paciente	Análisis de la adherencia del paciente a los tratamientos prescritos y comprender los factores que influyen en el cumplimiento del tratamiento.	Evaluar la relación entre la adherencia al tratamiento de la diabetes y los resultados de control glucémico en pacientes [17].
Estudios Economía de la Salud	Evaluar la relación costo-eficacia de medicamentos e intervenciones médicas en diversas poblaciones de pacientes.	Estudiar la eficacia de combinar EHR y Economía de la Salud para la gestión de enfermedades crónicas especialmente en pacientes con T2DM [18].
Asociaciones Genotipo-Fenotipo	Exploración de las relaciones entre variaciones genéticas y fenotipos observables en pacientes del mundo real para descubrir posibles determinantes genéticos de la respuesta a los medicamentos.	Estudios PheWAS para buscar asociaciones entre variantes genéticas y diferentes fenotipos [19].

Efectos de los Medicamentos a Largo Plazo	Estudiar los efectos a largo plazo de los medicamentos en los pacientes, la progresión de enfermedades y la calidad de vida utilizando la EHR.	Investigar la relación entre la exposición de medicamentos a largo plazo y efectos adversos en pacientes long-COVID [20].
---	--	---

Dada su capacidad para mejorar la atención clínica que ofrece la digitalización de las historias clínicas, la adopción de EHR se ha convertido en un aspecto crucial en los sistemas sanitarios modernos en todo el mundo. Sin embargo, esta adopción no es uniforme entre los países, y las tasas más altas se encuentran en aquellos con mayores recursos económicos (Figura 1). De hecho, estudios de mercado³ publicados en 2022 demostraron que el mercado de EHR está claramente dominado por América del Norte, el Reino Unido y los países de la Unión Europea (UE), mientras que la región de Asia-Pacífico (APAC) es la que experimenta el crecimiento más rápido. A pesar de que los EHR se definen como herramientas para mejorar la calidad de la atención médica, todavía existe una brecha significativa en la adopción de EHR en países con recursos limitados (LIC, del inglés *Low Income Countries*) [21]. Las deficiencias en la infraestructura, falta de compromiso por parte de las administraciones o empresas, problemas de interoperabilidad, escasez de apoyo y baja experiencia en tecnología de la información, se presentan como barreras importantes para implementar los sistemas EHR [22]. Sin embargo, en el año 2020, la pandemia global causada por la enfermedad del coronavirus (COVID-19) puso de manifiesto la necesidad de contar con sistemas avanzados de EHR y supuso un impulso para la adopción a nivel mundial [23].

³ <https://www.transparencymarketresearch.com/electronic-health-records-market.html>

Figura 1. Los 10 Países con Mayor Adopción de EHR en el Mundo



En el año 2021, la Comisión Europea llevó a cabo un estudio en el cual analizó la implementación de los registros médicos digitales en los países de la UE, Noruega y el Reino Unido, así como el nivel de interoperabilidad. Los resultados obtenidos de esta investigación revelaron un panorama heterogéneo en cuanto a la adopción y desarrollo de este sistema tecnológico. Esta diversidad se observó tanto entre naciones como dentro de diferentes regiones dentro de un mismo país [24]. La configuración de los sistemas de salud en Europa varía ampliamente: algunos son centralizados, otros descentralizados y pueden depender de aseguradoras tanto públicas como privadas, entre otras. Si a esta complejidad le añadimos consideraciones legales y de seguridad, nos encontramos en que las EHR rara vez son interoperables, ya que cada sistema por separado está diseñado para proteger la privacidad del paciente.

Esta falta de interoperabilidad tiene consecuencias que se extienden más allá de la atención médica e impacta en la investigación, ya que, al no poder acceder a conjuntos de datos más amplios y variados, se limita la capacidad de identificar tendencias y asociaciones estadísticas de manera efectiva. Para abordar esta problemática, los proveedores de EHR a menudo optan por combinar sus datos con otros conjuntos una vez que se han tomado medidas para garantizar el anonimato de los pacientes, cumpliendo con regulaciones de privacidad y el Reglamento General de Protección de Datos (GDPR, por sus siglas en inglés *General Data Protection Regulation*). En un intento por abordar estos desafíos, en el año 2017, la Agencia Europea del Medicamento (EMA, por las siglas en inglés *European Medication Agency*) presentó una iniciativa para la creación de un Modelo de Datos Comunes (CDM, por sus siglas en inglés *Common Data Model*) para toda la UE [25].

En el caso de España, el mismo informe [24] resaltó que existe una digitalización similar a la de los países más avanzados de Europa como Estonia, Dinamarca, Finlandia o Suecia, aunque se trata de un sistema altamente descentralizado en sus 17 comunidades autónomas [26]. Si bien algunas comunidades autónomas son líderes en Europa, la cobertura no abarca a todo el país y la interoperabilidad de los EHR entre estas comunidades es muy limitada. Esto tiene un impacto directo en la calidad de los servicios de salud ofrecidos a los ciudadanos, quienes podrían enfrentar dificultades al intentar acceder a sus datos médicos en diferentes regiones autónomas. Para ello, recientemente se ha presentado la *European Health Data Space* (EHDS) que servirá de entidad reguladora sobre la protección de datos así como el acceso de los ciudadanos a estos datos [27].

Cataluña es una de las comunidades con mayor digitalización y de los sistemas más avanzados de Europa. El Sistema Sanitario Público de Cataluña (SISCAT) se encarga desde el año 2000 de la gestión de datos y las tecnologías de la información para el empoderamiento del enfermo, la práctica asistencial, la gestión sanitaria y la asignación de recursos. Actualmente, el SISCAT está integrado por 68 centros hospitalarios, 434 centros de atención primaria, 102 centros de internamiento sociosanitarios, 40 centros de salud mental con internamiento, 428 ambulancias de soporte vital básico y avanzado y 4 helicópteros medicalizados del Sistema de Emergencias Médicas, así como más de 900 ambulancias de transporte sanitario no urgente (rehabilitación, diálisis, etc.) [28]. Mediante el acceso a la reutilización y cruce de los datos generados por el SISCAT, el Programa de Analítica de Datos para la Investigación y la Innovación en Salud (PADRIS) tiene la misión de poner a disposición de la comunidad científica los datos sanitarios relacionados para impulsar la investigación, la innovación y la evaluación en salud [29].

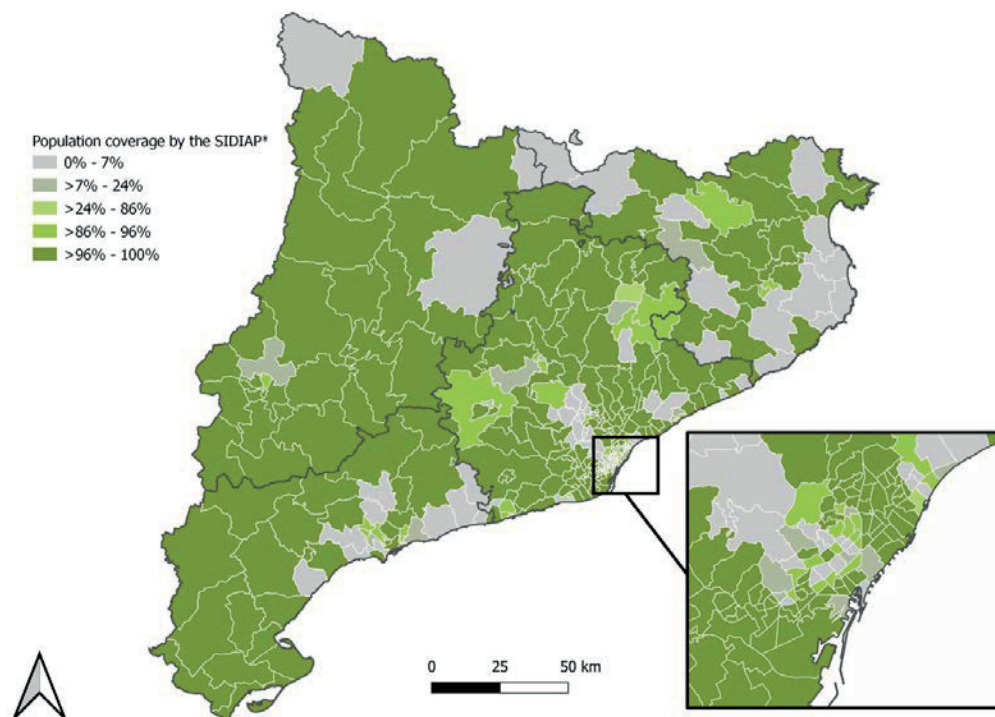
SISCAT colabora con diversos proveedores de servicios de salud, siendo el Instituto Catalán de la Salud (ICS) uno de los más destacados. El ICS, entidad pública adscrita al gobierno de la Generalitat de Cataluña, administra 283 equipos de atención primaria (EAP) y cuenta con alrededor de 8000 profesionales. Estos profesionales registran los datos de los pacientes en un sistema informático denominado Estación Clínica de Atención Primaria (eCAP), donde se almacenan las historias clínicas de unos 8 millones de pacientes, aproximadamente el 80% de la población de Cataluña.

A partir de las historias clínicas almacenadas en eCAP, el Sistema de Información de los Servicios de Atención Primaria (SISAP) ha desarrollado un sistema integral de explotación de datos. Su objetivo principal es proporcionar información relevante para la gestión clínica y generar indicadores significativos que beneficien a los profesionales de la atención médica [30].

De la alianza entre el SISAP del ICS y el Instituto de Investigación en Atención Primaria de Salud Jordi Gol (IDIAPJGol) nace, en enero de 2010, Sistema de Información para el Desarrollo de la Investigación en la Atención Primaria (SIDIAP⁴) [31],[32].

La plataforma SIDIAP incorpora los datos almacenados en eCAP desde 2006 y se ha establecido como una base de datos validada y de calidad en EHR. Su propósito es promover la investigación en el ICS y ser un referente en proyectos de investigación tanto a nivel nacional como internacional en el IDIAPJGol. SIDIAP incluye datos actualizados de los individuos en SISAP y a 30 de junio de 2021 engloba un total de 8,036,948 individuos en todo el territorio catalán (Figura 2), con un seguimiento longitudinal de 15.2 años (mediana, rango intercuartílico (IQR): 6.2 a 15.5). Del total, el 50.7% son mujeres y la mediana de edad es de 44 años (IQR: 25 a 60). [31].

Figura 2. Población Catalana Representada en SIDIAP



Martina Recalde et al., "The Information System for Research in Primary Care (SIDIAP)," *Int J Epidemiol*, vol 51, no. 6, pp e324–e336, Dec. 2022, doi:10.1093/ije/dyac068

⁴ Toda la información sobre SIDIAP en el enlace: <https://www.sidiap.org/index.php/es/>

1.2 Bases de Datos de EHR

En IDIAPJGol, la base de referencia de EHR es el SIDIAP, el sistema de bases de datos especialmente preparado para la investigación en atención primaria (PC, en inglés *Primary Care*). Los datos del SIDIAP se actualizan semestralmente, al final de cada año y a mediados del año natural, y están distribuidos por tablas según el tipo de información que contienen: datos demográficos, estado de salud, medicamentos, información clínica, resultados de laboratorio, entre otros (más detalle en Apéndice 1, p.111).

En SIDIAP, todas las tablas incluyen siempre:

- El identificador seudonimizado⁵ (idp) del paciente.
- La fecha completa del momento de entrada del registro.
- El código identificador de la variable, por ejemplo código diagnóstico de enfermedades CIE-10⁶ (en inglés, *ICD-10*).
- Un agrupador (AGR) que engloba diferentes códigos bajo una única etiqueta, definida por el equipo investigador o por el equipo de SIDIAP en base a los AGR empleados en otros proyectos de investigación.

Además, SIDIAP proporciona un catálogo que incluye la descripción de todos los códigos, los agrupadores AGR a los que corresponden y, en el caso de las variables categóricas⁷, su significado.

Tabla 3. Ejemplo de Registros Electrónicos de Diagnósticos en SIDIAP

idp ^a	dat	datfi ^b	cod ^c	AGR ^d
431asnroe3rfrdk8	10022010	27082010	J44.0	EPOC
63be208baqewdc	30112009	20112020	E11.1	T2DM

a. Identificador seudonimizado, b. Fecha inicio(dat) y final (datfi), c. Código ICD-10, d. Categoría AGR

⁵ Procedimiento que reemplaza el identificador real por uno o más identificadores artificiales o pseudónimos aptos para el análisis de datos.

⁶ Clasificación Internacional de Enfermedades, 10.^a edición definida por la Organización Mundial de la Salud, que determina la codificación de las enfermedades y una amplia variedad de signos, síntomas, entre otros.

⁷ Registros de observaciones que provienen de una característica cualitativa.

Aprobación Ética

La base de datos del SIDIAP cuenta con la aprobación del Instituto Catalán de la Salud (ICS), y sigue las directrices de la Declaración de Helsinki⁸ (Fortaleza, Brasil 2013). Cuyos principios para el uso de datos personales en investigación se centran en la minimización de estos datos, la integridad y confidencialidad (eliminando cualquier dato que pueda identificar directa o indirectamente al participante) y la transparencia (proporcionando información de los estudios que se realizan). Por lo tanto, no es necesario obtener el consentimiento de los participantes para su realización y los datos están seudonimizados. Además, los datos del SIDIAP se mantienen en un servidor seguro y se aplican agregaciones o supresiones en los registros de acuerdo con las regulaciones establecidas en el Reglamento General de Protección de Datos (GDPR) 2016/679 y la Ley Orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los Derechos Digitales.

No obstante, para acceder a los datos de SIDIAP, es necesario enviar el protocolo de estudio a evaluación y obtener la certificación por parte del Comité Ético de Investigación con Medicamentos (CEIm) del IDIAPJGol. Para realizar esta tesis, se necesitaron dos certificados:

1. Certificado para el desarrollo y validación del algoritmo. Código de referencia: AR20/029, SIDIAP 386 del 3 de junio de 2020.
2. Certificado para el análisis de la población con T2DM. Código de referencia: 21/026-P, del 24 de febrero de 2021.

1.3 Datos de Medicamentos

Para los medicamentos, SIDIAP cuenta con dos fuentes principales: las facturaciones en farmacias y las prescripciones electrónicas. Ambas fuentes codifican los medicamentos utilizando el Sistema de Clasificación Anatómica, Terapéutica, Química (código ATC-7, del inglés *Anatomical, Therapeutic, Chemical classification system*), que identifica el principio activo y el subgrupo terapéutico. Además, como se ha explicado anteriormente, SIDIAP incorpora un AGR para categorizar los códigos ATC-7 en grupos con características similares (por ejemplo, conjunto de antidiabéticos o anticoagulantes) para facilitar su manejo.

⁸ Declaración de la Asociación Médica Mundial (AMM) por los Principios Éticos Para Las Investigaciones Médicas En Seres Humanos

Prescripciones Electrónicas

Esta fuente comprende todas las prescripciones efectuadas por los profesionales de atención primaria (PC) a través del programa eCAP. SIDIAP pone a disposición los siguientes datos relativos a la prescripción electrónica:

- Código ATC-7 del medicamento.
- Fechas de inicio y finalización de la prescripción.
- Posología especificada en la receta, indicando las unidades por toma.
- Frecuencia de administración especificada por el profesional.
- Unidades contenidas en cada envase.
- Tipo de unidad proveedora de PC (UP⁹) que emite la receta.

Facturaciones en Farmacias

Desde 2006, SIDIAP dispone de todos los productos farmacéuticos financiados por el Sistema Nacional de Salud y dispensados mediante receta electrónica en las farmacias. Estos datos provienen de las facturaciones proporcionadas por las farmacias al CatSalut. La información sobre dispensaciones disponible en SIDIAP incluye:

- Código ATC-7 del medicamento.
- Fecha del mes de dispensación.
- Cantidad de envases retirados durante el mes.
- Unidades de medicamento contenidas en cada envase.
- Número de Dosis Diaria Definida (NDDD) de esa retirada.
- Coste total del medicamento cubierto por el CatSalut.
- Coste neto del medicamento para el paciente.

Visualización de los Datos de Medicamentos

Durante el desarrollo de la tesis, en colaboración con todas las expertas de la UEM, se pensó en un gráfico que representara visualmente los momentos de exposición de cada paciente a lo largo del tiempo (consulte la Figura 3). Este gráfico consta de dos ejes principales: vertical y

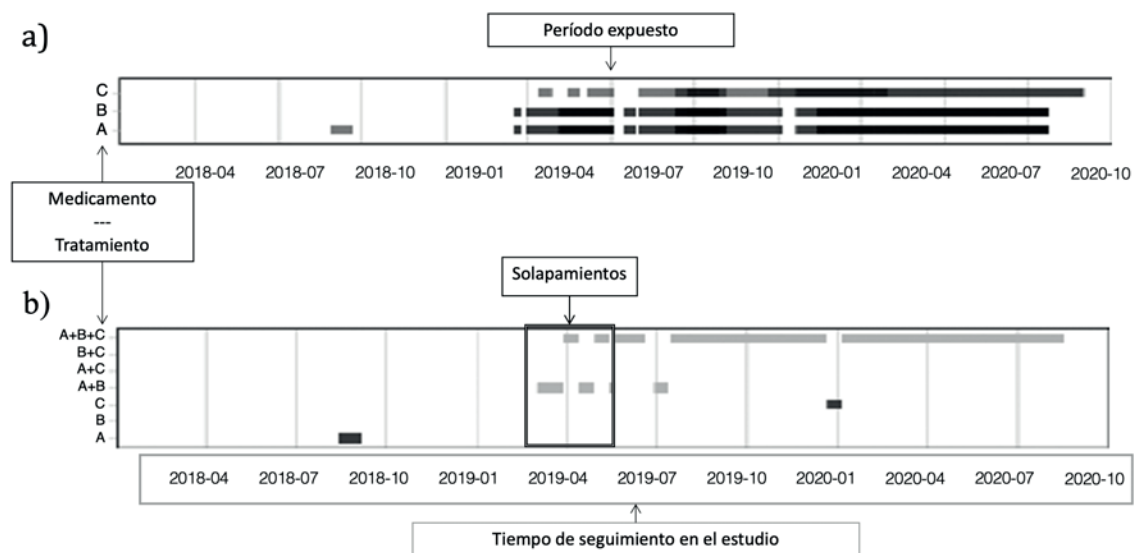
⁹ Estructura organizativa que dispone de recursos humanos, físicos y económicos para la prestación de servicios sanitarios de atención primaria.

horizontal. El eje vertical incluye los medicamentos (pueden ser agregadores o códigos ATC-7) que se desean representar, mientras que el eje horizontal abarca la ventana de tiempo de seguimiento del paciente, en días. Cada registro de medicamento se representa mediante segmentos o cajas en los cuales el extremo izquierdo del segmento indica el inicio del registro de prescripción o facturación electrónica, mientras que el cierre del segmento señala su finalización.

Este gráfico se usa para visualizar los medicamentos (“gráfico de datos crudos”) y para ilustrar los tratamientos (“gráfico de exposición a tratamientos”). La principal diferencia radica en el eje vertical: la de datos crudos presenta cada medicamento por separado, mientras que la segunda incluye los tratamientos junta en un mismo tratamiento todos los medicamentos que se registran simultáneamente en un mismo período de tiempo.

Este gráfico es útil para exponer patrones de tratamiento longitudinales y comportamiento en relación a otros eventos clínicos de interés. También se usa a lo largo de la tesis para ilustrar el funcionamiento del algoritmo “smooth”.

Figura 3. Representación Gráfica de los Datos de Medicamentos en EHR



Ejemplo simulado de un paciente al que se le prescriben tres medicamentos diferentes en el mismo período de tiempo. a) Gráfico de datos crudos. b) Gráfico del tratamiento al que el paciente está expuesto.

1.4 Analizar la Exposición a Medicamentos Usando Bases de Datos

En estudios con EHR, la información sobre la exposición a medicamentos suele obtenerse a través de la prescripción electrónica, facturación mensual de los medicamentos dispensados en las farmacias y de la receta electrónica, que sí incluye el día de dispensación en farmacia (por el momento este repositorio no está disponible en el SIDIAP). Esta información está ampliamente aceptada y utilizada en la investigación clínica, ya que la disponibilidad de datos registrados longitudinalmente permite una caracterización detallada tanto de la exposición a la medicación como para estudiar asociaciones o interacciones a eventos de interés. Con un correcto análisis, se puede generar una mayor evidencia de los efectos de los medicamentos en el mundo real, complementando lo que sabemos de los ensayos clínicos controlados aleatorizados [10].

Los tratamientos de estudio se clasifican según su duración en agudos o crónicos. Los tratamientos agudos pueden ser de corta duración (días hasta un máximo de 4 semanas), como los antibióticos o antiinflamatorios, o de duración intermedia (hasta 6 meses), como los antidepresivos o benzodiacepinas. Además, también pueden surgir tratamientos intermitentes, como el tratamiento agudo de la migraña con un patrón indefinido, u otros intermitentes con un patrón estacional, como el tratamiento para el trastorno por déficit de atención e hiperactividad en niños. Se considera que un tratamiento es crónico cuando perdura más de 6 meses y está dirigido a una patología para la cual no se espera su resolución ni la retirada del tratamiento.

1.5 Modelar el Tratamiento Longitudinal

En el contexto de la farmacoepidemiología o farmacovigilancia, uno de los principales objetivos es caracterizar o modelar el tratamiento en términos de duración [33], interrupciones o discontinuaciones [34], [35], adiciones, deleciones o cambios [36], y el cumplimiento en términos de adherencia a los tratamientos farmacológicos [17] [37]. El conjunto de cálculos que se realizan para lograr este nivel de detalle respecto al tratamiento de un paciente es lo que se conoce como modelización del tratamiento longitudinal. Aunque se ha avanzado en esta área para la monoterapia [38], esta modelización adquiere especial importancia al explorar la exposición a tratamientos de larga duración en condiciones como hipertensión, diabetes mellitus tipo 2 (T2DM, por sus siglas en inglés Type 2 Diabetis Mellitus) o enfermedad pulmonar obstructiva crónica (EPOC), el tratamiento a menudo cambia de monoterapia a terapia combinada (CT, en inglés *Combination Therapy*) de dos o más medicamentos [38], lo que aumenta significativamente el desafío de modelar

el tratamiento. En investigaciones previas realizadas en nuestro grupo de investigación en UEM [17], la CT en EHR suele generar patrones de tratamiento complejos que son difíciles de analizar o interpretar, ya que es difícil distinguir si son ocurrencias reales o errores de registro. Estos patrones pueden quedar ocultos a los investigadores o dar lugar a malentendidos, y para evitarlos a menudo se usan criterios poco homogéneos y estandarizados que pueden llevar a resultados sesgados y poco reproducibles.

Es para esta parte donde los algoritmos automáticos demuestran ser de gran utilidad para modelar los siguientes eventos (véase Figura 4):

Inicio del Tratamiento

El tratamiento inicial es el que se observa tras el diagnóstico de la enfermedad en estudio, en el caso de esta tesis, es la Diabetes Tipo 2 (T2DM). También se considera tratamiento inicial aquellos iniciados hasta 30 días antes de la fecha de diagnóstico. Este margen suele ser necesario debido a que las prescripciones no están ligadas a los diagnósticos y puede pasar que el registro de diagnóstico entre más tarde que la prescripción para ese problema de salud.

Monoterapia y Terapia Combinada

Para los análisis de ADM, la exposición al tratamiento se divide en fases de un solo medicamento, monoterapia, o si implica la exposición a ≥ 2 medicamentos, es decir, una terapia combinada (CT, por el inglés *Combination Therapy*).

Cambios de Tratamiento

Se considera un cambio de tratamiento, o “switch”, cuando el grupo de medicamentos que el paciente utiliza cambia a otro medicamento de un grupo o principio activo (PA) diferente. Este “switch” puede ser entre CT, entre monoterapias o puede implicar la transición de Monoterapia a CT, o viceversa.

Adición, Delección y Interrupción Temporal

En relación a las adiciones y delecciones, se considera una adición cuando se dispensa un nuevo PA o grupo de medicación mientras el medicamento previamente dispensado aún está en posesión del paciente. Delección, por otro lado, representa la sustracción de un medicamento o PA, manteniendo, por lo menos, uno al que ya se está expuesto. Las adiciones pueden llevar de monoterapia a CT o de CT a CT, pero nunca una adición conduce a monoterapia. Al contrario, una delección siempre parte de CT y puede cambiar a CT o a monoterapia. Es importante tener en cuenta que una adición o delección no se considera un “switch”.

Se define una interrupción temporal del tratamiento cuando existe un periodo sin administración de medicamentos entre dos periodos de exposición.

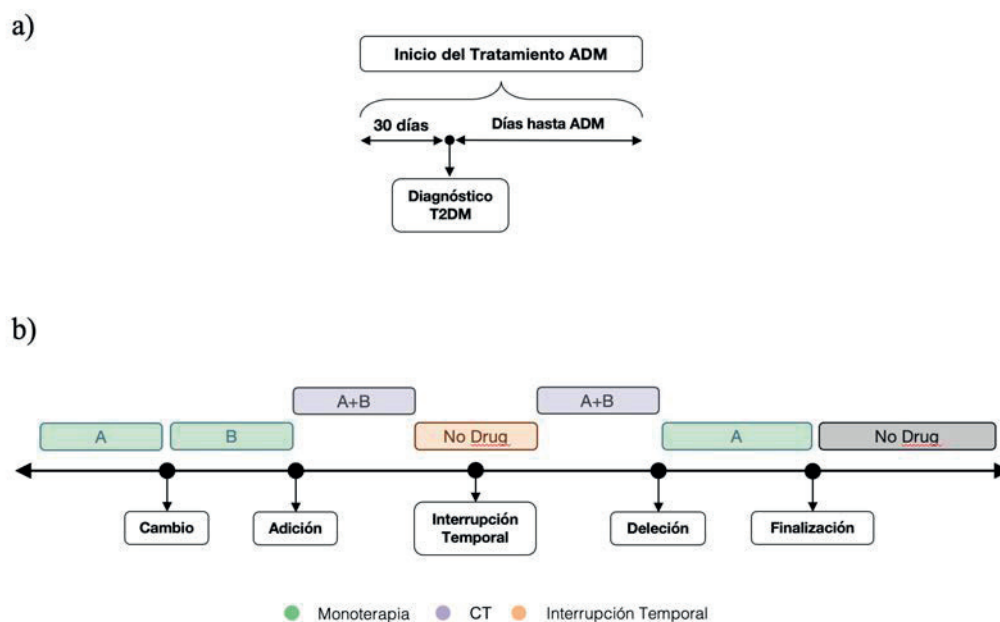
Finalización del Tratamiento

Este punto se alcanza cuando se concluye un tratamiento y, a partir de ese momento, no se registran más tratamientos durante el periodo de estudio.

Duración

Para este cálculo, el algoritmo “smooth” resulta especialmente útil. La duración se calcula como el número de días desde la fecha inicial hasta la fecha de cambio, adición, delección, interrupción o de finalización de un tratamiento. También se calcula el tiempo en monoterapia o en CT.

Figura 4. Eventos que Componen la Modelización Longitudinal del Tratamiento

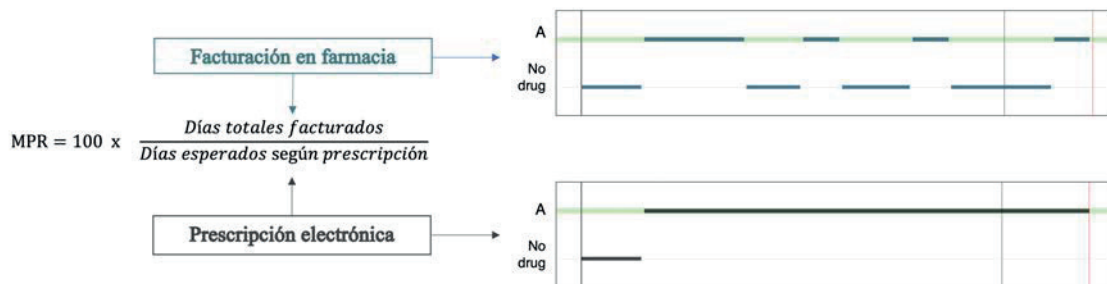


Panel a) representa el algoritmo que determina el inicio del tratamiento ADM. La parte b) es un esquema con los diferentes eventos de interés en farmacoepidemiología. El código de colores diferencia entre monoterapia, terapia combinada (CT) e interrupción temporal.

1.6 Calcular el Cumplimiento

La adherencia se calcula como una aproximación de la tasa de posesión del medicamento (MPR, por las siglas en inglés *Medication Possession Ratio*). Es la proporción de los días en posesión del medicamento por parte del paciente (registros de dispensación en la farmacia) en relación con el número total de días esperados (prescripciones electrónicas) que dura ese tratamiento (véase Figura 5).

Figura 5. Cálculo del Cumplimiento



A partir del MPR, se clasifican a los pacientes como adherentes o no adherentes en función de si superan o no el umbral del 80%. Tanto el MPR como la adherencia se puede calcular en monoterapia como en terapias combinadas.

1.7 Tratamiento de la Diabetes Tipo 2

Antidiabéticos

Los medicamentos antidiabéticos (ADM, por su siglas en inglés *Antidiabetic Medication*) son fundamentales en el tratamiento y manejo de la diabetes. Están especialmente diseñados para regular los niveles de glucosa en sangre en personas que padecen diabetes tipo 2 (T2DM, más detalle sobre esta población en Apéndice 1, p.115), contribuyen a mejorar la sensibilidad a la insulina, disminuir la producción de glucosa por parte del hígado y potenciar la utilización de glucosa por parte de las células del organismo.

Existen varios tipos de medicamentos antidiabéticos, cada uno con su propio mecanismo de acción y efectos en el cuerpo. A parte de las insulinas, en esta tesis estudiamos los principales grupos de ADM (véase Tabla 4 para más detalle):

- **Metformina:** Es uno de los medicamentos más comúnmente recetados a los pacientes con T2DM. Contribuye a reducir la producción de glucosa por parte del hígado y mejora la sensibilidad a la insulina.
- **Sulfonilureas:** Estos medicamentos estimulan al páncreas a liberar más insulina, lo que ayuda a disminuir los niveles de glucosa en sangre.
- **Inhibidores de la DPP-4 (DPP4i):** Actúan aumentando los niveles de incretinas, hormonas que estimulan la liberación de insulina y reducen la producción de glucosa.
- **Inhibidores del SGLT2 (SGLT2i):** Estos medicamentos permiten que los riñones eliminen el exceso de glucosa a través de la orina, contribuyendo a reducir los niveles de glucosa en sangre.
- **Análogos del GLP-1 (GLP1):** Son administrados mediante inyecciones y funcionan estimulando la liberación de insulina y reduciendo la producción de glucosa.

Estos medicamentos pueden ser recetados individualmente o en combinación para alcanzar un mejor control de la diabetes. La elección del tratamiento dependerá de varios factores, como las necesidades específicas de cada paciente, la gravedad de la enfermedad y la presencia de otras condiciones de salud.

Tabla 4. Descripción (en inglés) de Códigos ATC-7 y Grupos de Antidiabéticos (ADM)

ANTIDIABÉTICOS (ADM)	DESCRIPCIÓN	CÓDIGOS ATC-7
Metformin	Metformin	A10BA02
	Gliclazide	A10BB09
	Glimepiride	A10BB12
	Glibenclamide	A10BB01
	Glipizide	A10BB07
Sulfonylurea		

	Chlorpropamide	A10BB02
	Glyquidone	A10BB08
	Glisentide	A10BB91
	Liraglutide	A10BJ02
	Dulaglutide	A10BJ05
	Semaglutide	A10BJ06
	Exenatide	A10BJ01
GLP1	Lixisenatide	A10BJ03
	Sitagliptin	A10BH01
	Linagliptin	A10BH05
	Vildagliptin	A10BH02
	Alogliptin	A10BH04
DPP4i	Saxagliptin	A10BH03
	Insulin glargine	A10AE04
	Insulin (human)	A10AC01
	Insulin aspartate	A10AB05
	Insulin (human)	A10AB01
	Insulin detemir	A10AE05
	Insulin lispro	A10AB04
	Insulin aspartate	A10AD05
Insulin	Insulin glulisine	A10AB06

	Insulin lispro	A10AD04
	Insulin (human)	A10AD01
	Insulin degludec	A10AE06
	Insulin lispro protamine	A10AC04
SGLT2i	Empagliflozin	A10BK03
	Dapagliflozin	A10BK01
	Canagliflozin	A10BK02
	Ertugliflozin	A10BK04

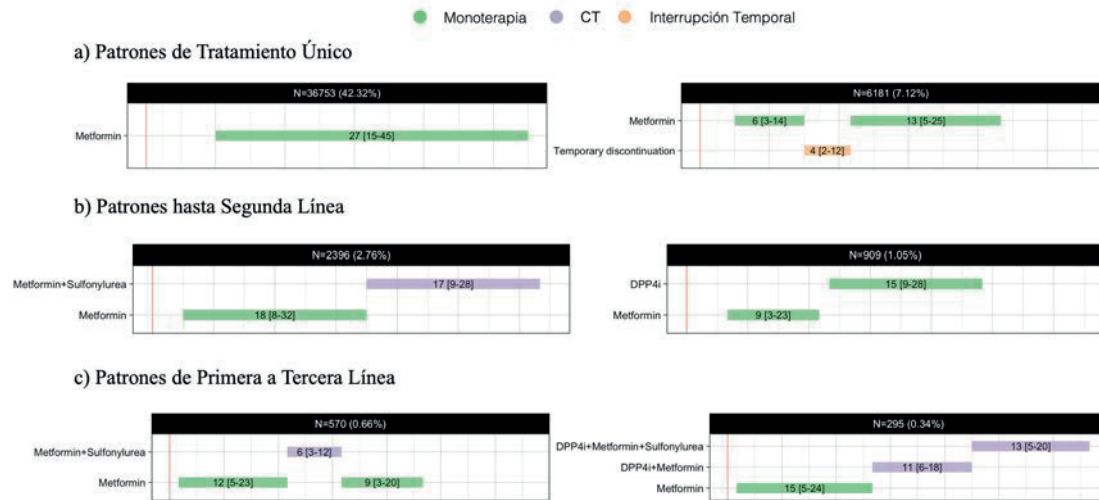
Patrones Longitudinales de Tratamiento Antidiabético

A partir de modelar el tratamiento longitudinal de cada paciente, se pueden calcular patrones longitudinales de tratamientos con ADM. Estos patrones no son más que secuencias que se presentan sistemáticamente en conjuntos de pacientes. Los patrones, al no ser producto del azar, suelen estar vinculados con otros eventos de interés. En el caso de la presente tesis, en el primer estudio se analiza la asociación de los patrones longitudinales con los cambios en los niveles de HbA1c observados tres meses antes de la aparición de un cambio, adición, delección o interrupción del tratamiento. En el segundo estudio, se relaciona los patrones de tratamiento a la reducción de la HbA1c por debajo del 8% en pacientes con tratamiento previo de Metformina a los que se le añade un ADM de segunda línea.

La Figura 6 incluye seis de estos patrones longitudinales detectados en pacientes con tratamiento antidiabético. En el panel a) se incluyen dos ejemplos que pasan por un único tratamiento. El de la derecha muestra un intervalo de interrupción temporal (en color naranja) entre dos períodos de exposición con Metformina. En el panel b), se presentan patrones con dos líneas de tratamiento. A la izquierda, se observan los pacientes que inician con Metformina y, al cabo de 18 meses, cambian a CT de Metformina con Sulfonilurea (adición). Mientras que en el de la derecha, se presenta un “switch” de Metformina a DPP4i en monoterapia pasados 9 meses. Finalmente, en c), hay dos patrones que abarcan de primera a tercera línea de tratamiento. El de la izquierda es un ejemplo de adición de Metformina a CT con Sulfonilurea, que luego se retira

para volver a Metformina en monoterapia (delección). En la derecha, se muestra un grupo de pacientes con doble adición. Comienzan en monoterapia con Metformina, se agrega DPP4i a los 15 meses y mantienen CT durante 11 meses para, finalmente, comenzar una tercera línea con un tratamiento triple.

Figura 6. Patrones Longitudinales en ADM



Patrones longitudinales según el número de cambios y si inician con monoterapia o CT. La amplitud del segmento representa la duración, en mediana e IQR entre paréntesis.

Datos de Hemoglobina Glicada

La hemoglobina glicada (HbA1c) es un análisis de sangre que proporciona información acerca de los niveles promedio de glucosa o azúcar en la sangre durante los últimos tres meses. Esta prueba se utiliza tanto para diagnosticar la T2DM como la prediabetes y también es un análisis fundamental para el seguimiento de la enfermedad. El resultado de la prueba se expresa en forma de porcentaje, a mayor nivel porcentual, mayores han sido los niveles de glucosa en la sangre [39]:

- Un nivel de HbA1c normal se encuentra por debajo del 5.7%.
- La prediabetes oscila entre el 5.7 y el 6.4%. La prediabetes es un factor de riesgo para el desarrollo de la T2DM.
- Valores iguales o superiores a 6.5% indican diabetes.

En Cataluña, se han establecido pautas de armonización para el tratamiento con ADM [40]. Estas pautas recomiendan que las personas con T2DM se realicen una prueba de HbA1c al menos dos veces al año. Esto permite que los profesionales de la salud, basándose en los resultados obtenidos, establezcan los objetivos del tratamiento, adapten la terapia y supervisen el control de la diabetes.

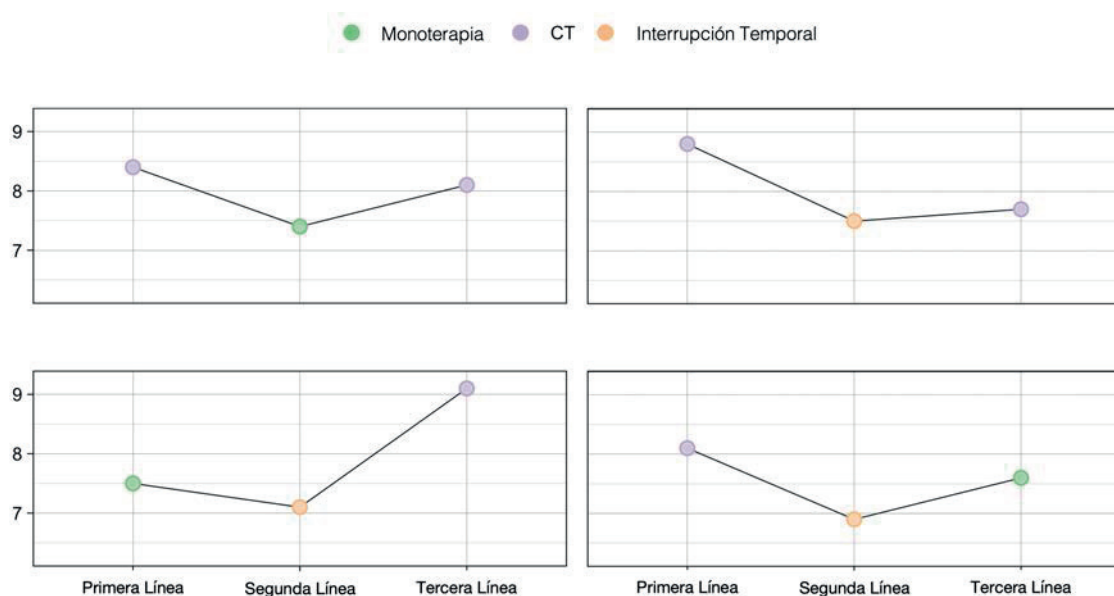
Uso en Estudios Observacionales

En SIDIAP, las pruebas de HbA1c se registran en las tablas de resultados de laboratorio, utilizando el porcentaje como su unidad de medida.

Las directrices médicas recomiendan realizar estas pruebas cada seis meses. Así, la HbA1c se convierte en una variable ampliamente utilizada en estudios de EHR en pacientes con T2DM, ya que está disponible para la mayoría de los pacientes, presenta un buen seguimiento a lo largo del tiempo y es un dato con nivel de calidad alto.

En concreto, la HbA1c se suele emplear para analizar la efectividad del tratamiento y optimizar las estrategias de tratamiento (véase Figura 7), identificar posibles factores de riesgo para hipertensión o dislipidemia, investigar la relación entre los niveles de HbA1c y la incidencia de complicaciones a largo plazo, como enfermedad cardiovascular, neuropatía o nefropatía diabética.

Figura 7. Asociación de Hba1c con los Cambios en la Pauta de Tratamiento

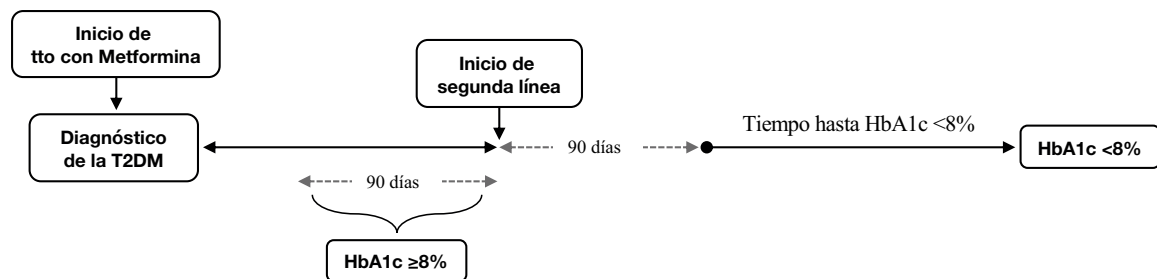


Incluye pacientes que han experimentado hasta tres líneas de tratamiento con ADM (eje horizontal). Los puntos en el gráfico representan el valor mediano de las mediciones de HbA1c tomadas tres meses

antes de un cambio en el tratamiento. Los colores utilizados en el gráfico indican si el cambio fue hacia monoterapia, CT o si hubo una interrupción temporal del tratamiento.

Para capturar los niveles de HbA1c en las EHR, se suelen establecer ventanas temporales de 90 o 180 días. Estas ventanas se definen antes y después de los eventos de interés y sirven tanto para no perder información relevante (en caso que las ventanas sean muy estrechas) cómo para dar un tiempo para evaluar el efecto de este resultado con otros eventos de interés (ejemplo en Figura 8).

Figura 8. Esquema de Utilización de HbA1c en el Estudio de Pacientes con T2DM



Diseño para estudiar el efecto de los tratamientos de segunda línea en la HbA1c para analizar la evolución de la enfermedad.

1.8 La Influencia de la Calidad de los Datos

Las historias clínicas en su forma actual suelen estar compuestas por conjuntos de datos complejos, sin estructura y no siempre perfectos. Un análisis sistemático publicado en 2020 sugiere que las dimensiones de calidad de los datos y los métodos utilizados en las EHR no son uniformes en la literatura, lo que dificulta las evaluaciones de calidad debido a la naturaleza compleja y heterogénea de estos datos [41].

Podemos definir hasta cinco atributos o dimensiones de interés para evaluar la calidad de las EHR tanto para su uso rutinario como para la investigación [42]:

- Exhaustividad: Registros completos, disponibles y de fácil acceso.
- Precisión: Garantizar que los elementos que componen la EHR sean correctos y estén libres de errores.
- Concordancia: Asegurar consistencia en los datos del paciente entre elementos dentro de la misma EHR o en relación con otras fuentes de datos.

- Coherencia: Validar que los datos sean coherentes y tengan sentido en relación con otros elementos del sistema EHR.
- Actualidad: Asegurarse de que la información disponible sobre el paciente sea suficiente para reflejar su estado de salud en todo momento.

Alcanzar los niveles de calidad deseada en las EHR representa una labor multidisciplinaria y de amplio alcance, que demanda considerables recursos humanos y esfuerzo económico. En todos los ámbitos aún existe margen de mejora, ya sea en motivar o educar al profesional de la salud (o el paciente) que introduce los datos, el mantenimiento y actualización de la infraestructura tecnológica que los almacena, la profesionalización del equipo técnico que los gestiona y procesa, o en la mejora de los métodos que emplean los propios usuarios finales para plantearlos, analizarlos y reportar los resultados [43]. Este posiblemente sea uno de los obstáculos o desafíos más significativos que enfrentan actualmente las EHR.

Los grupos de investigación forman parte de los usuarios finales mencionados anteriormente y no son ajenos a estos desafíos, ya que a la hora de analizar las EHR deben enfrentar una serie de dificultades que afectan la calidad y la coherencia de los resultados:

- La variabilidad de los datos entre diferentes proveedores de atención médica y sistemas de registro puede resultar en datos inconsistentes y difíciles de comparar.
- La falta de estandarización en la terminología médica y los códigos de diagnóstico pueden llevar a problemas de codificación y etiquetado.
- Además, la inclusión o exclusión de ciertos pacientes basada en criterios específicos puede introducir sesgos y afectar la representatividad de la muestra.

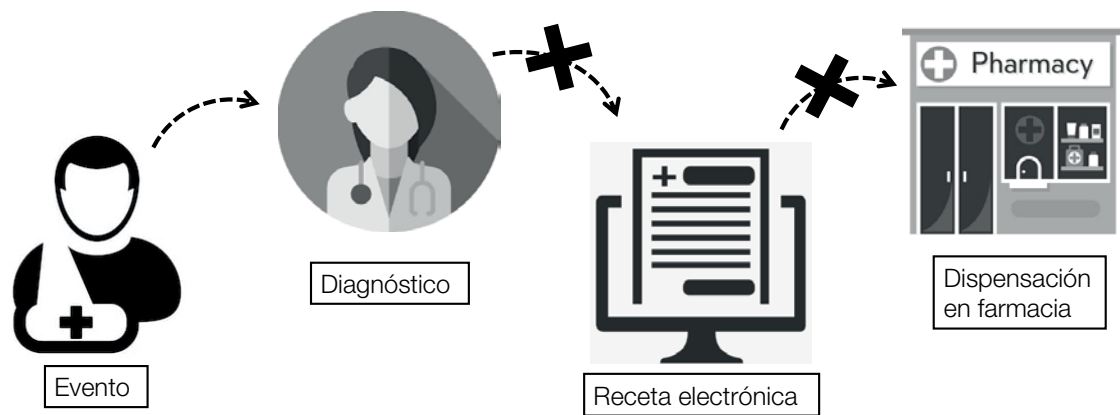
1.9 Limitaciones de los Datos de Medicamentos en SIDIAP

Los datos de SIDIAP se caracterizan por su calidad y por estar especialmente orientada hacia la investigación. Sin embargo, al igual que sucede con todas las bases de datos de EHR, también presenta ciertas limitaciones [31]. Centrándonos en los datos relacionados con medicamentos, es fundamental considerar dos desafíos principales.

El primero es que los registros de diagnóstico, prescripciones electrónicas y facturaciones de farmacia son registrados de forma independientemente y no han de estar vinculadas entre sí. Un evento puede llevar a un diagnóstico pero no es obligatorio que en el momento de su registro se incluya una prescripción electrónica. De manera similar, la prescripción y la dispensación en la farmacia no están necesariamente vinculadas entre sí. Como resultado, en SIDIAP no se dispone

de una secuencia completa que abarque desde el momento en que el paciente experimenta un evento y busca atención en AP hasta que va a la farmacia a por el medicamento recetado (Figura 9). Además, la concordancia temporal entre los registros no está garantizada. De hecho, la prescripción electrónica proporciona la fecha de inicio con precisión en día, mes y año, mientras que en la facturación solo se dispone del mes y año de la dispensación.

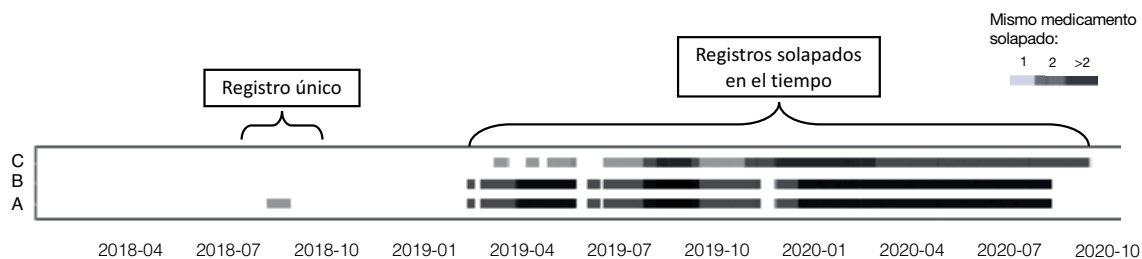
Figura 9. Discrepancias Entre la Realidad y los Registro Electrónicos en SIDIAP



Secuencia esperada que sigue un paciente en contraposición a los registros electrónicos en la base de datos de SIDIAP.

La segunda limitación se relaciona con las prescripciones electrónicas, donde el sistema eCAP registra todas las entradas de prescripciones realizadas por el profesional. Si hay alguna errónea o se cancela la prescripción, esta sigue presente en el sistema generando multitud de errores de registro. Además, existe un problema con la duración de las prescripciones. A cada prescripción se le asigna automáticamente una fecha de finalización, la cual puede ser modificada por el profesional pero a menudo se mantiene la automáticamente asignada. Esto resulta en la presencia de solapamientos entre prescripciones del mismo medicamento, especialmente prevalente cuando analizamos tratamientos de larga duración (Figura 10).

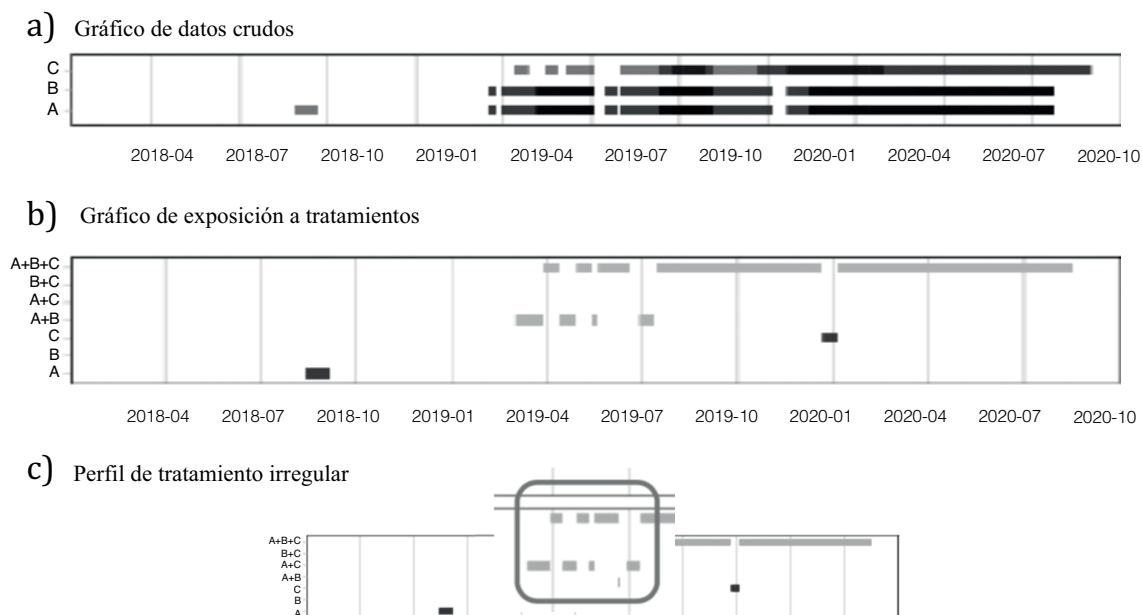
Figura 10. Solapamientos en las Prescripciones Electrónicas



La escala de grises indica el número de prescripciones del mismo tipo solapados en el tiempo.

Para afrontar estas dificultades, los grupos de investigación deben limpiar los registros para mejorar la calidad. Establecen criterios o algoritmos basados en premisas que no siempre son objetivas y que tienen limitada aplicabilidad fuera del contexto específico del estudio en cuestión. Además, en estudios farmacoepidemiológicos donde a menudo se exploran tratamientos combinados, la complejidad aumenta significativamente ya que a menudo aparecen perfiles en el tratamiento con cambios constantes (irregulares) que son de difícil explicación (véase Figura 11).

Figura 11. Tratamientos Combinados y Perfiles Irregulares

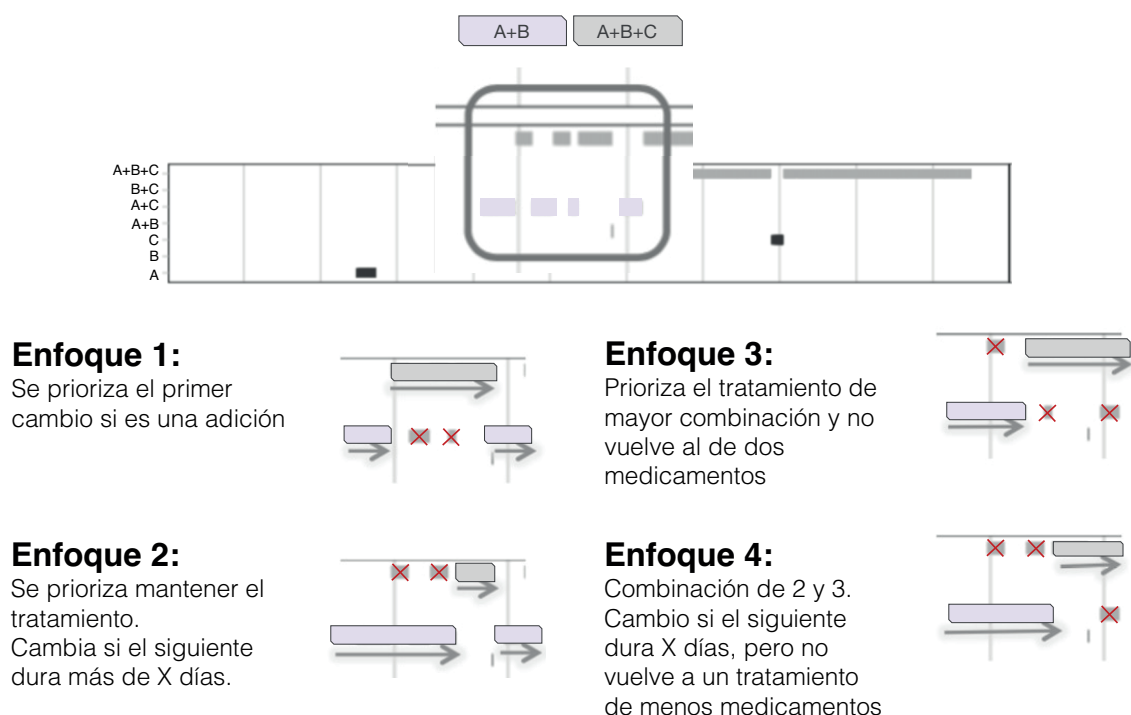


Se ilustra el mismo ejemplo de la figura 10. En la parte a), se presentan los datos crudos e individuales de tres medicamentos para una misma patología. En la parte b), se muestra el perfil de exposición a

estos medicamentos y el momento en que dos o más medicamentos se solapan, formando así un tratamiento combinado. En la parte c), se resalta un momento de la exposición en el que el perfil presenta irregularidades.

Frente a esta situación, nuevamente se hace necesario establecer criterios que eliminen o den sentido clínico a estos perfiles irregulares, y estos suelen ser muy específicos. Por ejemplo, en la Figura 12 se presentan cuatro posibles enfoques para abordar el perfil irregular de la Figura 11. En el primer enfoque, se prioriza el cambio a un tratamiento con tres medicamentos; en el segundo, se prioriza la 'estabilidad', lo que implica que no se acepta el cambio hasta que el nuevo tratamiento se mantenga constante en el tiempo. El tercer enfoque vuelve a priorizar el tratamiento combinado, haciéndolo más prolongado, ya que no considera que el paciente regrese a una terapia de dos medicamentos. Finalmente, el cuarto enfoque es una combinación de los criterios establecidos en los escenarios dos y tres, prioriza cierta constancia en el tiempo y, además, no permite que el tratamiento regrese a la terapia de dos medicamentos.

Figura 12. Formas de Afrontar los Perfiles de Tratamiento Irregulares



La figura presenta hasta 4 enfoques diferentes que varían el resultado según los criterios establecidos.

En este contexto se entiende que existe una necesidad de desarrollar métodos como el algoritmo “smooth”, central de esta tesis, como una alternativa automática y generalizable para abordar estas complicaciones.

Tabla 5. Otras Dificultades de Analizar las EHR Para Obtener Evidencias Científicas

El análisis de EHR presenta ciertas dificultades particulares que se deben tener en consideración para obtener evidencias fiables:

- **Sesgo de Selección:** Los EHR tienden a incluir principalmente a pacientes que han necesitado atención médica. Esto afecta la generalización de los resultados a la población en general.
- **Sesgo de Información:** Los datos faltantes o incorrectos pueden introducir sesgos de información en los análisis, lo que puede afectar la validez de los resultados.
- **Sesgo de Supervivencia:** Los pacientes que abandonan el sistema de atención médica pueden no estar representados y afectar la interpretación de los resultados.
- **Dimensiones Elevadas y Multicolinealidad:** Puede dificultar la identificación de relaciones significativas entre variables.
- **Ruido y Variabilidad:** Debido a diferentes prácticas de registro, medición y entrada de datos.
- **Variables de Confusión:** Identificar y controlar estas variables es esencial para evitar conclusiones erróneas.
- **Muestreo Desbalanceado:** Pueden estar desbalanceados en términos de distribución de clases o grupos, lo que puede afectar la capacidad de detectar patrones y relaciones en los análisis.
- **Validación Externa Limitada:** Puede ser desafiante, ya que los datos son específicos de ciertos sistemas de atención médica y no siempre son fácilmente comparables con otros conjuntos de datos.

Estas dificultades requieren una buena planificación del análisis, de métodos estadísticos apropiados y experiencia en técnicas de control de calidad de datos.

Para afrontarlos, los grupos de investigación suelen establecer criterios durante la selección de la población y la preparación de los datos, implementan métodos para estandarizar la terminología y los códigos, y aplican técnicas para manejar los datos faltantes. Por ello, es necesario formar grupos de corte transversal que incluya la colaboración entre expertos en informática médica, estadísticos y gestores de datos para garantizar la calidad y la integridad de los resultados obtenidos. En estudios de farmacoepidemiología con EHR, este aspecto cobra particular relevancia y existen iniciativas como la de ISPOR¹⁰ cuyo grupo de trabajo tiene como objetivo establecer directrices de buenas prácticas para mejorar la validación y la reproducibilidad de los estudios con estas bases de datos [44].

¹⁰ ISPOR Taskforce for the Use of EHR for Health Technology Assessment.

<https://www.ispor.org/member-groups/task-forces/use-of-electronic-health-records-for-health-technology-assessment>

2. Hipótesis

Los estudios farmacoepidemiológicos basados en EHR requieren el desarrollo de métodos que garanticen la obtención de resultados confiables y reproducibles respecto al análisis de los tratamientos. En la actualidad, estos estudios a menudo se basan en criterios válidos, pero poco aplicables a diferentes poblaciones o a contextos de uso de medicamentos diversos. La creación de un algoritmo automatizado permitiría simplificar y estandarizar el procesamiento de datos y mejorar la calidad y la reproducibilidad de los resultados obtenidos.

3. Objetivos

El propósito central de esta tesis doctoral es desarrollar un algoritmo automático destinado a identificar el tratamiento más probable al que un paciente ha estado expuesto durante el periodo de seguimiento y las modificaciones temporales que se observan durante el seguimiento de los tratamientos.

De manera más detallada, los objetivos de este estudio de doctorado son los siguientes:

1. Diseñar y desarrollar un algoritmo para mejorar la calidad y reproducibilidad en los análisis con medicamentos utilizando EHRs.
2. Utilizar el algoritmo para modelar la exposición a tratamientos crónicos, ya sea en solitario o en combinación.
3. Emplear el algoritmo para mejorar los análisis de cumplimiento (adherencia) a tratamientos simples o combinados.
4. Analizar patrones y perfiles longitudinales de tratamientos en pacientes con patología crónica, identificando los cambios, adiciones o retiradas.

En la Tabla 6 se presenta la relación entre las publicaciones y los objetivos de esta tesis.

Tabla 6. Relación entre Objetivos de la Tesis Doctoral y los Trabajos Publicados

ARTÍCULO DE INVESTIGACIÓN	OBJETIVO			
	1	2	3	4
Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study	✓	✓		
Longitudinal treatment patterns in patients recently diagnosed with type 2 diabetes mellitus in Catalonia		✓	✓	✓
Impact of Second-Line Combination Treatment for Type 2 Diabetes Mellitus on Disease Control: A Population-Based Cohort		✓	✓	✓

4. Compendio de Artículos

4.1 JMIR Medical Informatics

El primer artículo incluye todo el desarrollo del algoritmo “smooth”, su implementación en registros de salud y la validación con expertos independientes, primer objetivo de la tesis (descrito en Tabla 6).

Artículo titulado *"Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study"*

Publicado en *JMIR Medical Informatics*

vol. 10, no. 11, p. e37976, Nov. 2022,

DOI: [10.2196/37976](https://doi.org/10.2196/37976)

Factor de Impacto en 2023: 3.2 (Q2)

[Original Paper](#)

Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study

Dan Ouchi^{1,2}, MSc; Maria Giner-Soriano^{1,2,3}, PhD; Ainhoa Gómez-Lumbreras^{1,4}, PhD; Cristina Vedia Urgell^{1,2,3}, MD; Ferran Torres⁵, PhD; Rosa Morros^{1,2,3,6}, PhD

¹Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina, Barcelona, Spain

²Facultat de Medicina. Departament de Farmacologia, Toxicologia i Terapèutica, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain

³Unitat de Farmàcia. Servei d'Atenció Primària Barcelonès Nord i Maresme, Institut Català de la Salut, Barcelona, Spain

⁴Department of Pharmacotherapy, College of Pharmacy, University of Utah, Salt Lake City, UT, United States

⁵Unitat de Bioestadística Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain

⁶Spanish Clinical Research Network Platform, Barcelona, Spain

Corresponding Author:

Dan Ouchi, MSc

Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina

Gran Via de les Corts Catalanes 587, àtic

Barcelona, 08007

Spain

Phone: 34 934824110

Email: douchi@idiapjgol.info

Abstract

Background: Since the use of electronic health records (EHRs) in an automated way, pharmacovigilance or pharmacoepidemiology studies have been used to characterize the therapy using different algorithms. Although progress has been made in this area for monotherapy, with combinations of 2 or more drugs the challenge to characterize the treatment increases significantly, and more research is needed.

Objective: The goal of the research was to develop and describe a novel algorithm that automatically returns the most likely therapy of one drug or combinations of 2 or more drugs over time.

Methods: We used the Information System for Research in Primary Care as our reference EHR platform for the smooth algorithm development. The algorithm was inspired by statistical methods based on moving averages and depends on a parameter Wt , a flexible window that determines the level of smoothing. The effect of Wt was evaluated in a simulation study on the same data set with different window lengths. To understand the algorithm performance in a clinical or pharmacological perspective, we conducted a validation study. We designed 4 pharmacological scenarios and asked 4 independent professionals to compare a traditional method against the smooth algorithm. Data from the simulation and validation studies were then analyzed.

Results: The Wt parameter had an impact over the raw data. As we increased the window length, more patient were modified and the number of smoothed patients augmented, although we rarely observed changes of more than 5% of the total data. In the validation study, significant differences were obtained in the performance of the smooth algorithm over the traditional method. These differences were consistent across pharmacological scenarios.

Conclusions: The smooth algorithm is an automated approach that standardizes, simplifies, and improves data processing in drug exposition studies using EHRs. This algorithm can be generalized to almost any pharmacological medication and model the drug exposure to facilitate the detection of treatment switches, discontinuations, and terminations throughout the study period.

(JMIR Med Inform 2022;10(11):e37976) doi: [10.2196/37976](https://doi.org/10.2196/37976)

KEYWORDS

electronic health records; data mining; complex drug patterns; algorithms; drug utilization; polypharmacy; EHR; medication; drug combination; therapy; automation; drug exposition; treatment; adherence

<https://medinform.jmir.org/2022/11/e37976>

JMIR Med Inform 2022 | vol. 10 | iss. 11 | e37976 | p. 1
(page number not for citation purposes)

Introduction

The recent rise in the use of electronic health records (EHRs) has had a major impact on epidemiological research. These databases provide a low-cost means of accessing longitudinal data such as demographic, vital signs, administrative, medical and pharmacy claims, clinical, and patient-centered data on large populations for epidemiologic research [1,2].

However, in their current form, EHRs are complex and imperfect data sets that can be enhanced in a dizzying number of often ineffective ways. Although the challenges of working with EHRs in clinical trials have been identified [3-5], more research is needed to develop new and better ways to use them.

From the data mining perspective, addressing these data is particularly challenging as the outcomes can be significantly affected depending on the quality, validity, completeness, and heterogeneity of the available data [6]. Besides technical perspective, researchers have their particular ways of addressing EHR, dealing with EHR complexity, and their decisions have been shown to have a significant impact to the results [7,8]. Approaching these challenges in a heterogeneous and biased way favors the emergence of inconsistencies between similar studies [9].

In studies with EHR-based databases, information on drug exposure is usually obtained from electronic prescription, electronic dispensation, or invoice of drugs. This information is widely used and accepted in clinical research as the availability of longitudinally recorded data allows for a detailed characterization of both the exposure to medication and the outcome of interest, and mining the data contained within EHRs can potentially generate a greater understanding of medication effects in the real world, complementing what we know from randomized control trials [10].

Focusing on pharmacovigilance or pharmacoepidemiology when using EHRs, one of the main objectives is to characterize the therapy in terms of duration [11], discontinuation [12,13], changes [14], and adherence to pharmacological treatments [15]. Although progress has been made in this area for monotherapy [16], when we study treatment exposure in diseases such as hypertension, diabetes, or chronic obstructive pulmonary disease, treatment often switches from monotherapy to combinations of two or more drugs, which significantly increases the challenge of characterizing the treatment. In our experience [17], polytherapy in EHR-based studies creates

complex treatment patterns that are challenging to analyze or interpret, can be blinded to researchers, and can be a source of misunderstanding as it is difficult to distinguish whether they are real occurrences or recording errors. To address this, we propose a novel algorithm called *smooth* to obtain the most likely therapy of one or more drugs over time.

Methods

Data Sources

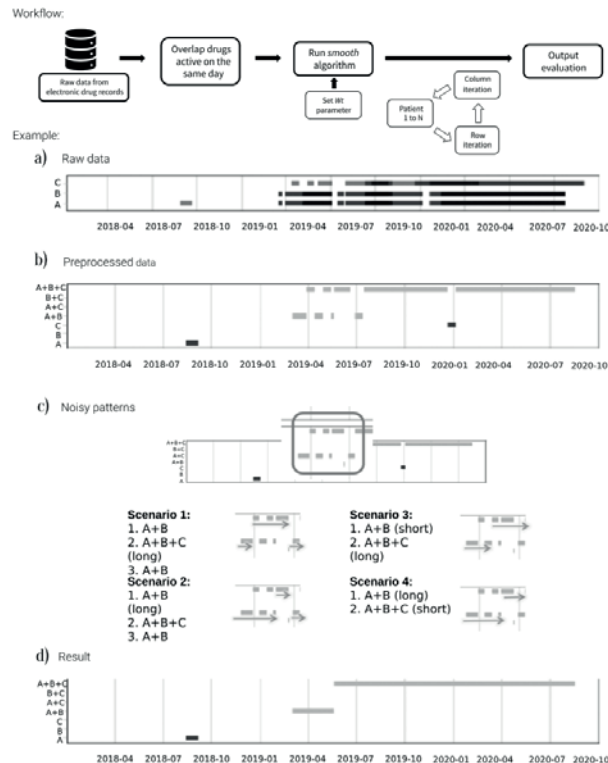
We used the Information System for Research in Primary Care (SIDIAP) [18] as our reference EHR platform for the algorithm development. The SIDIAP includes information recorded by health professionals during routine visits at 287 primary health care centers from the Catalan Health Institute (Institut Català de la Salut).

The platform includes information on disease diagnoses (*International Classification for Diseases, 10th Edition*), drug prescriptions and drug invoices in the primary care setting (Anatomical Therapeutic Chemical [ATC] classification system), and clinically relevant parameters (eg, weight, blood pressure, laboratory tests) as well as sociodemographic characteristics. It is also linked to a hospital discharge database for patients admitted to the Catalan Health Institute hospitals (30% of the SIDIAP population). The SIDIAP has pseudonymized records for more than seven million people and is representative of the Catalan population in terms of age, sex, and geographic distribution [19].

For the algorithm development and validation study, we used a subset of patients drawn from all Catalan Health Institute primary care centers. From this population, we also obtained sociodemographic characteristics: sex, age, country of origin, profession, socioeconomic index, smoking habits, alcohol intake, institutionalization in nursing homes, comorbidities, and electronic prescriptions of pharmacological treatments.

In SIDIAP, electronic prescriptions and drug invoices are stored in longitudinal format. Each record comprises the pseudonymized patient identifier, ATC code, and prescription or invoice date. The end of the prescription is determined by the health professional, whereas in drug invoice records we only have the month in which the invoice was made, and thus the end of the treatment is usually inferred based on the number of packages collected. Each prescription or invoice is recorded independently of the health problem, and duplicate records or overlaps are common (Figure 1A).

Figure 1. Overview of the smooth algorithm workflow with an illustrated example. (A) Horizontal axis is follow-up time in days, and the vertical axis is the different drugs prescribed. Box length indicates the period in which the prescription is active. (B) Patient profile after combining all active prescriptions in the same day, the first step of data process. (C) Example of complex patterns and 4 ways to overcome them. (D) Result obtained after passing the data through the smooth algorithm.



Ethics Approval

The study protocol was approved by the Research Ethics Committee of Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol; AR20/029, SIDIAP 386 on June 3, 2020). This is a database research study that has been conducted according to the guidelines of the Declaration of Helsinki (Fortaleza, Brazil 2013) and does not require consent from the people included to participate or for publication. The need for consent was waived by the Research Ethics Committee of IDIAPJGol as it is deemed unnecessary according to European legislation (Regulation [EU] 2016/679).

Finding the Most Likely Therapy Using the Smooth Algorithm

The process has 2 parts: data mining to look for treatments recorded daily and applying the smooth algorithm to the data (Figure 1).

For the first part, we looked at all drugs of interest that are active on the same day (Figure 1A and Figure 1B). This step simplifies

the prescription or invoice records but frequently reveals complex patterns that should be considered before conducting any analysis. Data are processed based on assumptions about those patterns by individual researchers and therefore give nonhomogeneous results. In Figure 1C, we imagined 4 scenarios (but there could be more) to handle the same problem (highlighted area during the first half of 2019). While some researchers may consider that the first therapy lasts until we observe a change in the treatment, others with different backgrounds may decide that a change from double to triple therapy can only be considered if the triple therapy lasts longer than an arbitrary period (eg, 60 days).

The smooth algorithm is inspired by statistical methods using simple moving averages that calculate trends or smooth time series [20]. For EHRs, we changed the concept of moving averages to a moving window from where we choose the most frequent treatment. Thus, by moving the window one day at a time, we identify the most frequent pattern over the study period (Figure 1D).

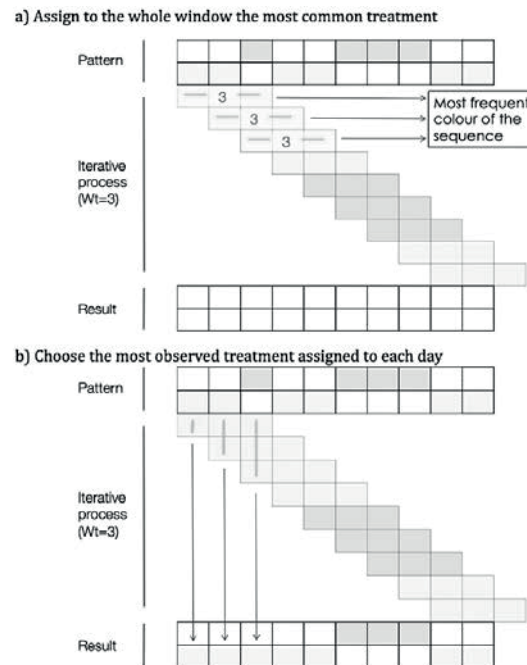
In Figure 2, we can see a detailed description of the algorithm. It is an iterative process that works as follows:

<https://medinform.jmir.org/2022/11/e37976>

JMIR Med Inform 2022 | vol. 10 | iss. 11 | e37976 | p. 3
(page number not for citation purposes)

- Starting from the first prescription or dispensation at day t_i , we opened a window of specific length in days (Wt) in which we search for the most frequent treatment
- That treatment was assigned to the whole window unless we had a draw (2 treatments are active for the same number of days), in which case we carried the previous treatment (t_{i-1}) forward.
- We then shifted the window forward one day at a time, repeating the process until the end of follow-up (Figure 2A).
- After the first iteration, we had up to Wt possible treatments (or candidates) for each day
- Finally, we chose the candidate most frequently observed on that day (Figure 2B).

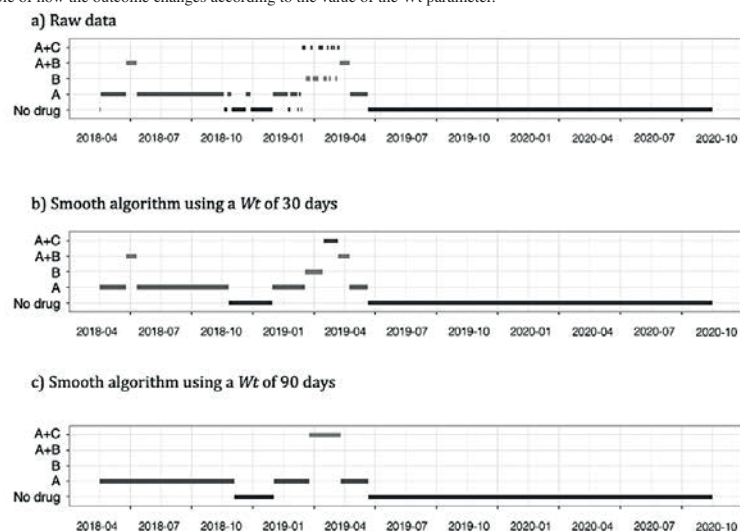
Figure 2. Smooth algorithm in detail.



Window Size

The length of Wt is the only parameter that needs to be defined beforehand, and its value can modify the outcome (Figure 3). The length of Wt determines the level of smoothing, and the value can go from 1 day to the total days of follow-up. Thus, for Wt of 1 day, we are not changing the data while for a Wt equal to the number of days of follow-up, we expect to reduce all records to the most frequent treatment. Therefore, small values of the Wt parameter will not significantly modify the raw data, whereas increasing the size of the window is expected to have a larger impact on the data.

In a more in-depth analysis, a simulation study was conducted using 7132 patients who were under long-term administration of aspirin, statins, beta-blockers, and angiotensin-converting enzyme (ACE) inhibitors or angiotensin-receptor blockers between 2018 and 2020. The smooth algorithm was run on this data using 6 Wt values: 10, 20, 30, 45, 60, and 90 days. For each Wt , we counted the number of patients with at least one change in the treatment pattern; out of these, we calculated the percentage of smooth as the ratio of the number of days changed by the algorithm divided by the total number of days with active treatment.

Figure 3. Example of how the outcome changes according to the value of the Wt parameter.

Validation Study

To understand how the algorithm performs from a clinical or pharmacological perspective, we conducted a validation study. We identified 4 pharmacological scenarios where the algorithm could be needed—combination of 3 or more drugs (platelet aggregation inhibitors, beta-blockers, and ACE inhibitors), treatments likely to discontinue (antidepressants), long-term combination of 2 drugs (insulins and oral antidiabetics), and short-term treatment (systemic antibiotics)—and asked to 4 independent professionals with experience in databases and drug exposure studies to compare a traditional method with the smooth algorithm.

The traditional method is a more intuitive and simple approach, commonly observed in the literature, to address noise and variability in electronic drug records [21,22]. Briefly, it starts with the first treatment observed and accepts a treatment change only if the new one is longer than a certain period of time. This period is generally arbitrary, an assumption done by the researcher based on the characteristics of the drug. For our validation study, we set the period to 60 days except for antibiotics (the short-term treatment), with a period of 15 days.

For the smooth algorithm, the Wt parameter was set to 60 days in all 4 scenarios.

Before conducting the validation, we prepared a training session with the 4 reviewers consisting in an introduction to the data, drugs of study (including the selected ATC codes), explanation of the algorithms, and discussion of the common criteria to apply during the validation. From their feedback, and after the training, we decided to include all health problems related to the treatment as it may facilitate the evaluation and give more importance to clinical criteria (see validated sample in Multimedia Appendix 1, Figure S1).

Our primary objective was changes to the original data; we analyzed whether the algorithms improved, worsened, or made no changes to the original data. In addition, we asked the reviewers to choose between the traditional method and the smooth algorithm and evaluate its value for detecting treatment switches and/or discontinuations.

Each professional reviewed 100 patient records with one-quarter of the records being assigned to all reviewers to analyze consistency across validations. To reduce potential biases, reviewers were blinded and they did not know which method or algorithm generated the results (Table 1).

Table 1. Description of drugs and distribution of samples in the validation study.

Treatment pattern of use	Description (ATC ^a code)	Prescriptions in the data set	Samples analyzed, total (per reviewer)	Samples repeated across reviewers, total (per reviewer)
Short-term drugs	Systemic antibiotics (J01)	10,846,282	80 (20)	40 (10)
Likely to discontinue	Antidepressants (N06a)	3,859,496	80 (20)	40 (10)
Long-term combinations of 2 drugs	Insulins and oral antidiabetics (A10)	22,271,154	120 (30)	60 (15)
Combination of 3 or more drugs	Platelet aggregation inhibitors (B01Ac)	21,253,742	120 (30)	60 (15)

^aATC: Anatomical Therapeutic Chemical.

Statistical Methods

We determined that a sample size of 400 patients would be enough to ensure 80% power assuming a minimum effect size of 0.3062 with two degrees of freedom for a Chi-square test under a significance level of 5%.

Categorical variables were described with relative and absolute frequencies, and results from numerical variables were reported using means and standard deviations. For the validation study, we used the Chi-square test to evaluate differences between the traditional method and smooth algorithm on performance compared with the raw data. The algorithm was programmed in R (version 4.1.0, R Foundation for Statistical Computing), and all analyses were performed in R.

Results

Impact of W_t Parameter and Simulation Study

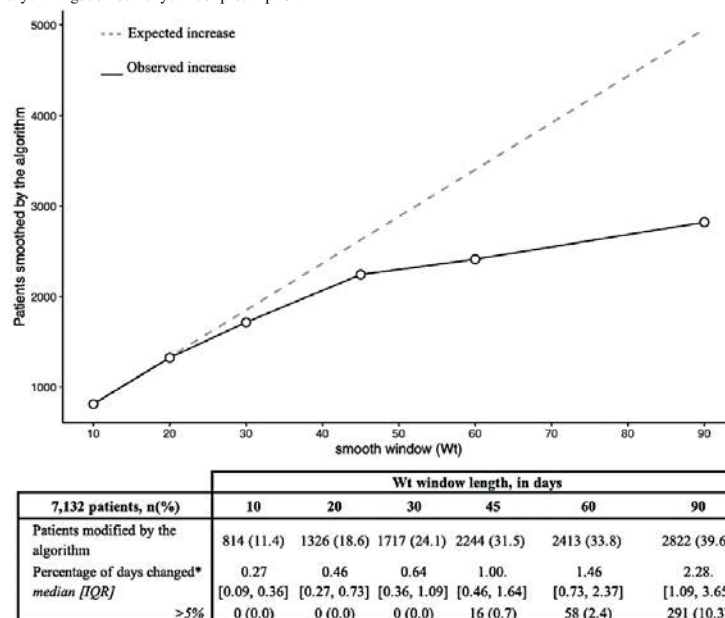
In Figure 3, we show how the results changed according to the W_t value. In the example, for the raw data we note that during the first period of follow-up, the main treatment was a monotherapy followed by a complex pattern during the first 6-month semester of 2019 (Figure 3A). To reduce noise in the pattern, we applied the smooth algorithm using W_t values of 30 and 90 days. With a W_t of 30 days (Figure 3B), we retained the combination of 2 drugs during the early stages of follow-up; during months with more changes, patient moved from

monotherapy to a combination of 2 drugs (A+C and A+B) prior to switching to the A monotherapy. In contrast, with a W_t of 90 days (Figure 3C), the entire treatment pattern was simplified. During the first year of follow-up, we observed a monotherapy; during the most complex pattern, the algorithm smoothed the changes to a single combination of A+C before returning to a monotherapy.

Results of the simulation study are represented in Figure 4. With a W_t of 10 days, 11.4% (814/7132) had their treatment pattern changed, while with a 90-day window, 39.6% (2822/7132) had their treatment pattern modified; 31.5% (2244/7132), 33.8% (2413/7132), and 39.6% (2822/7132) of patients with W_t values of 45, 60, and 90 days, respectively, saw at least 1 change. Thus, the effect of W_t was not linear as the expected progression of the number of patients being smoothed was different than the results from the simulation.

As a relative measure, we reported the percentage of days changed, ranging from 0.27 (IQR 0.09, 0.36) to 2.28 (IQR 1.09, 3.65). Thus, for each of the W_t values 10, 20, 30, 45, 60, and 90 days in a 1000-day period of follow-up, the algorithm modified 2.7, 4.6, 6.4, 10, 14.6, and 22.8 days, respectively. At windows from 10 to 30 days, the percentage of days changed always remain below the 5%, but as we increased the W_t to 45, 60, and 90 days, we started to observe patients with more than 5% of the data smoothed.

Figure 4. Statistics of the simulation study using 6 W_t values on 7132 patients under treatment for cardiovascular disease between 2018 and 2020. *100 x Number of days changed / Total days under prescription.



Validation Study

Multimedia Appendix 2 includes the results of the validation study. For the 400 samples, the smooth algorithm improved the raw data for 56.8% (227/400) of individuals, while 42.5% (170/400) benefited from using the traditional method. In 39% (156/400) of the samples, the outcome provided by each algorithm did not change the patterns, and 4.2% (17/400) of cases reported worsening after being processed by the smooth algorithm. The traditional method resulted in a worse outcome for 18.5% (74/400) of the samples, and the observed differences between algorithms were statistically significant ($P < .001$).

Significant differences were also observed between algorithms stratified by scenario. With a combination of 3 or more drugs (platelet aggregation inhibitors, beta-blocker, and ACE inhibitors), drugs likely to be discontinued (antidepressants), and the combination of 2 or more long-term drugs (insulins and antidiabetics), the smooth algorithm improved 69.2% (83/120), 61.3% (49/80), and 60.0% (72/120) of samples, respectively. In short-term treatment (systemic antibiotics), 90.0% (72/80) of samples did not show changes and 28.7% (23/80) were improved by the smooth algorithm.

As for the samples validated by the 4 professionals (Table 1), they decided in 88.0% (44/50) of patients to choose the smooth algorithm over the traditional method (Multimedia Appendix 2 and Multimedia Appendix 1, Figure S2). It was in the short-term treatment scenario where we observed less of a consensus, 70% (7/10), whereas for the rest of scenarios the 4 reviewers agreed in 93.3% (14/15), 100% (10/10), and 86.7% (13/15) of the samples, respectively.

The smooth algorithm performed better than the traditional method in detecting discontinuations (350/366, 95.6%) and treatment switches (138/230, 60.0%; see Multimedia Appendix 1, Table S1).

Discussion

Principal Findings

Since the use of EHR databases began in pharmacoepidemiologic studies, researchers have been trying to establish algorithms to model drug exposure [23]. This becomes even more challenging when trying to assess drug exposures with multiple pharmacologic treatments, which happens quite often in older people, as they are prescribed with up to 5 drugs simultaneously [24]. Thus, we have developed an automatic algorithm to model drug exposure through EHRs, which standardizes the data mining process to obtain more consistent and replicable results across studies.

The algorithm is inspired by time series forecasting methods and requires a parameter to be set beforehand. This is commonly observed in similar statistical methods such as autoregressive models or moving averages [25], and it is known that the value of the parameter can modify the outcome significantly [26-28]. The simulation study shows the impact of the W_t value. Small values hardly change the original data, but as the parameter value increases, the raw data can be affected to the point of losing clinical relevance. In the worst-case scenario, we

observed changes in up to 40% of the patients, with 75% of those having at least 1% of the records smoothed.

Interestingly, the simulation shows that at a certain W_t value, the number of individuals modified reaches a plateau. The data changed by the algorithm are less than expected, particularly when W_t is greater than 30 days, suggesting that independently of the parameter, some patients will never be changed by the algorithm.

In the validation study, we observed that most times our algorithm improved the data patterns. It was designed to improve polypharmacy exposure assessment, and we were interested in the results for combinations of 3 or more drugs. In this scenario, both the traditional and smooth approaches demonstrated usefulness, and the percentages of improved samples were similar, although the smooth algorithm performed significantly better. These differences were also observed in the other scenarios, and we believe that the smooth algorithm not only improves the treatment pattern but also does not make it worse. In addition, the performance was not affected by the window length, since for antibiotics and antidepressants (short- and long-term drug use, respectively) the smooth algorithm performed well using the same W_t window.

The traditional method proved to be a good approach, and similar versions are being used in other studies [21,22,29], but it differs according to drug or study characteristics and so is less generalizable. In fact, it has not worked well for systemic antibiotics even though we specifically changed it to fit for its characteristics. Overall, the validations for the smooth algorithm were consistent between scenarios and reviewers.

Potential Uses and Strengths

We believe that the smooth algorithm has significant potential to assess exposure for treatment combinations, especially in chronic treatments, since it allows us to have a time sequence of exposure to the treatment. This sequence allows us to better model the drug exposition and detect discontinuations, switches, and periods of interaction with other drugs of short duration. In addition, it can be of great help in estimating adherence to a combination treatment [30]. With the smooth algorithm, we can easily calculate the exposure time using only electronic prescriptions without the need to know the dosage posology.

From a clinical point of view, the smooth algorithm has great advantage when estimating polypharmacy adherence [31]. Patients affected by chronic conditions in need of polypharmacy may have differing levels of adherence to individual medications within their regimen, and this could lead to varying health outcomes and misleading results if the methodological approach assumes as equivalent adherence to all medications. These patients also face other acute conditions requiring the addition of drugs or modification of doses while maintaining their actual medication regimen.

Another research area in which our algorithm could be of utility is the study of adverse events due to lack of effectivity (antibiotics or hypertension treatments) and drug-drug interactions (anticoagulants and nonsteroidal anti-inflammatories combined for short time periods).

Although we cannot ever know a patient's true adherence, the smooth algorithm is an automated way to analyze EHR data offering methodological consistency across studies. In several studies, assumptions were made prior to the data processing, and this may have impacted the results of the analyses by introducing bias on the final results [32]. Algorithms like smooth can help standardize these assumptions and minimize inconsistencies between studies with similar databases.

Limitations

Due to the nature of the algorithm eliminating complex patterns, we may lose relevant exposures of a short period of time, and smooth is not recommended for all scenarios. Similarly, smooth may not work well in long-acting drugs.

We were not able to capture the posology with the SIDIAP database, so we could not estimate the length of treatment through the dose prescribed. The treatment doses can change throughout the year (decreased use of diuretics during summertime, when traveling, etc).

From a technical perspective, using the smooth algorithm is a time-consuming process. To run the algorithm on big data sets like EHRs, good information technology is needed. The time needed to complete the process may vary depending on the number of patients and follow-up time.

In addition, before running the algorithm, we must set a parameter, W_t , that allows us to choose between precision and simplicity. Setting this parameter is not straightforward, and it

is important to understand the effect on the outcome to use a good value. This is an inherent limitation of the algorithm and, as a guide, we recommend setting the W_t value within the range of 30 to 60 days to reduce complex patterns without compromising relevant information. Moreover, the W_t can be changed so the algorithm can be used in short (antibiotics) and long-term (antidepressants) treatments as well as in drug combinations for chronic conditions such as diabetes and hypertension.

Another limitation of the study is that the validation was done with the traditional method instead of other algorithms as a comparator. Moreover, in our experience, we commonly see the traditional method being used with some differences or criteria according to the framework or objectives of the study. For example, in projects with the European Medicines Agency, we have never seen a method or an automatic approach to deal with drug exposure other than the one we call traditional [33].

Conclusion

The smooth algorithm is an automated approach to estimate the most likely drug exposure pattern. We proved that it standardizes, simplifies, and improves the data processing steps before performing the study analysis; can model the drug exposure to detect cotreatment, switches, discontinuations, and treatment terminations; and facilitates adherence calculations throughout the study period. In future pharmacoepidemiological studies, we aim to further validate the algorithm and analyze the impact the algorithm can have on the main results.

Authors' Contributions

DO designed and developed the algorithm and performed the statistical analyses. MG and DO collected the data. All authors contributed to the validation study. DO, AG, MG, FT, and RM were involved in the interpretation of the data. DO wrote the manuscript and designed the tables and figures. All authors critically revised the report and approved the final version to be submitted for publication. The corresponding author confirms that he had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary material.

[\[DOCX File , 1022 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Results of the validation study.

[\[DOCX File , 15 KB-Multimedia Appendix 2\]](#)

References

1. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health* 2016 Mar;37:61-81. [doi: [10.1146/annurev-publhealth-032315-021353](#)] [Medline: [26667605](#)]
2. Rogers JR, Lee J, Zhou Z, Cheung YK, Hripesak G, Weng C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. *J Am Med Inform Assoc* 2021 Jan 15;28(1):144-154. [doi: [10.1093/jamia/ocaa224](#)] [Medline: [33164065](#)]
3. Cowie MR, Blomster JJ, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017 Jan;106(1):1-9 [FREE Full text] [doi: [10.1007/s00392-016-1025-6](#)] [Medline: [27557678](#)]

4. Lu Z. Information technology in pharmacovigilance: benefits, challenges, and future directions from industry perspectives. *Drug Healthc Patient Saf* 2009 Oct;1:35-45 [FREE Full text] [doi: [10.2147/dhps.s7180](https://doi.org/10.2147/dhps.s7180)] [Medline: [21701609](https://pubmed.ncbi.nlm.nih.gov/21701609/)]
5. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. *Risk Manag Healthc Policy* 2011 May;4:47-55 [FREE Full text] [doi: [10.2147/RMHP.S12985](https://doi.org/10.2147/RMHP.S12985)] [Medline: [22312227](https://pubmed.ncbi.nlm.nih.gov/22312227/)]
6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: [10.1136/amiajnl-2011-000681](https://doi.org/10.1136/amiajnl-2011-000681)] [Medline: [22733976](https://pubmed.ncbi.nlm.nih.gov/22733976/)]
7. Salas M, Lopes LC, Godman B, Truter I, Hartzema AG, Wettermark B, et al. Challenges facing drug utilization research in the Latin American region. *Pharmacoepidemiol Drug Saf* 2020 Nov 17;29(11):1353-1363. [doi: [10.1002/pds.4989](https://doi.org/10.1002/pds.4989)] [Medline: [32419226](https://pubmed.ncbi.nlm.nih.gov/32419226/)]
8. Rudin RS, Friedberg MW, Shekelle P, Shah N, Bates DW. Getting value from electronic health records: research needed to improve practice. *Ann Intern Med* 2020 Jun 02;172(11_Supplement):S130-S136. [doi: [10.7326/m19-0878](https://doi.org/10.7326/m19-0878)]
9. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, Joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Pharmacoepidemiol Drug Saf* 2017 Sep 15;26(9):1018-1032 [FREE Full text] [doi: [10.1002/pds.4295](https://doi.org/10.1002/pds.4295)] [Medline: [28913963](https://pubmed.ncbi.nlm.nih.gov/28913963/)]
10. Yao L, Zhang Y, Li Y, Sanseau P, Agarwal P. Electronic health records: implications for drug discovery. *Drug Discov Today* 2011 Jul;16(13-14):594-599. [doi: [10.1016/j.drudis.2011.05.009](https://doi.org/10.1016/j.drudis.2011.05.009)] [Medline: [21624499](https://pubmed.ncbi.nlm.nih.gov/21624499/)]
11. Størring H, Pottegård A, Hallas J. Determining prescription durations based on the parametric waiting time distribution. *Pharmacoepidemiol Drug Saf* 2016 Dec 26;25(12):1451-1459. [doi: [10.1002/pds.4114](https://doi.org/10.1002/pds.4114)] [Medline: [27670969](https://pubmed.ncbi.nlm.nih.gov/27670969/)]
12. Lyu H, Zhao S, Yoshida K, Tedeschi S, Xu C, Nigwekar S, et al. Delayed denosumab injections and bone mineral density response: an electronic health record-based study. *J Clin Endocrinol Metab* 2020 May 01;105(5):1 [FREE Full text] [doi: [10.1210/clinem/dgz321](https://doi.org/10.1210/clinem/dgz321)] [Medline: [31894244](https://pubmed.ncbi.nlm.nih.gov/31894244/)]
13. Mascarenhas J, Mehra M, He J, Potluri R, Lofgren C. Patient characteristics and outcomes after ruxolitinib discontinuation in patients with myelofibrosis. *J Med Econ* 2020 Jul 31;23(7):721-727 [FREE Full text] [doi: [10.1080/13696998.2020.1741381](https://doi.org/10.1080/13696998.2020.1741381)] [Medline: [32159402](https://pubmed.ncbi.nlm.nih.gov/32159402/)]
14. Sharman J, Kabadi SM, Clark J, Andorsky D. Treatment patterns and outcomes among mantle cell lymphoma patients treated with ibrutinib in the United States: a retrospective electronic medical record database and chart review study. *Br J Haematol* 2021 Feb 23;192(4):737-746. [doi: [10.1111/bjh.16922](https://doi.org/10.1111/bjh.16922)] [Medline: [33095453](https://pubmed.ncbi.nlm.nih.gov/33095453/)]
15. Li X, Cole SR, Westreich D, Brookhart MA. Primary non-adherence and the new-user design. *Pharmacoepidemiol Drug Saf* 2018 Apr 19;27(4):361-364 [FREE Full text] [doi: [10.1002/pds.4403](https://doi.org/10.1002/pds.4403)] [Medline: [29460385](https://pubmed.ncbi.nlm.nih.gov/29460385/)]
16. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* 2019 Nov;394(10211):1816-1826. [doi: [10.1016/s0140-6736\(19\)32317-7](https://doi.org/10.1016/s0140-6736(19)32317-7)]
17. Giner-Soriano M, Sotorra Figuerola G, Cortés J, Pera Pujadas H, Garcia-Sangenis A, Morros R. Impact of medication adherence on mortality and cardiovascular morbidity: protocol for a population-based cohort study. *JMIR Res Protoc* 2018 Mar 09;7(3):e73 [FREE Full text] [doi: [10.2196/resprot.8121](https://doi.org/10.2196/resprot.8121)] [Medline: [29523501](https://pubmed.ncbi.nlm.nih.gov/29523501/)]
18. Bolibar B, Fina AF, Morros R, Garcia-Gil MDM, Hermosilla E, Ramos R, et al. [SIDIA database: electronic clinical records in primary care as a source of information for epidemiologic research]. *Med Clin (Barc)* 2012 May 19;138(14):617-621. [doi: [10.1016/j.medcli.2012.01.020](https://doi.org/10.1016/j.medcli.2012.01.020)] [Medline: [22444996](https://pubmed.ncbi.nlm.nih.gov/22444996/)]
19. García-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, Ramos R, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIA). *Inform Prim Care* 2011 Jun 01;19(3):135-145 [FREE Full text] [doi: [10.14236/jhi.v19i3.806](https://doi.org/10.14236/jhi.v19i3.806)] [Medline: [22688222](https://pubmed.ncbi.nlm.nih.gov/22688222/)]
20. Watson GS. Smooth regression analysis. *Sankhyā Indian J Stat* 1964;Series A(26):359-372.
21. Xu H, Aldrich M, Chen Q, Liu H, Peterson N, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *J Am Med Inform Assoc* 2015 Jan;22(1):179-191 [FREE Full text] [doi: [10.1136/amiajnl-2014-002649](https://doi.org/10.1136/amiajnl-2014-002649)] [Medline: [25053577](https://pubmed.ncbi.nlm.nih.gov/25053577/)]
22. Hughey J, Colby J. Discovering cross-reactivity in urine drug screening immunoassays through large-scale analysis of electronic health records. *Clin Chem* 2019 Dec;65(12):1522-1531 [FREE Full text] [doi: [10.1373/clinchem.2019.305409](https://doi.org/10.1373/clinchem.2019.305409)] [Medline: [31578215](https://pubmed.ncbi.nlm.nih.gov/31578215/)]
23. Andrade SE, Kahler KH, Frech F, Chan KA. Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiol Drug Saf* 2006 Aug;15(8):565-574. [doi: [10.1002/pds.1230](https://doi.org/10.1002/pds.1230)] [Medline: [16514590](https://pubmed.ncbi.nlm.nih.gov/16514590/)]
24. Pazan F, Wehling M. Polypharmacy in older adults: a narrative review of definitions, epidemiology and consequences. *Eur Geriatr Med* 2021 Jun 10;12(3):443-452 [FREE Full text] [doi: [10.1007/s41999-021-00479-3](https://doi.org/10.1007/s41999-021-00479-3)] [Medline: [33694123](https://pubmed.ncbi.nlm.nih.gov/33694123/)]
25. Hannan EJ, Rissanen J. Recursive estimation of mixed autoregressive-moving average order. *Biometrika* 1983;70(1):303. [doi: [10.1093/biomet/70.1.303-b](https://doi.org/10.1093/biomet/70.1.303-b)]
26. Durbin J. Efficient estimation of parameters in moving-average models. *Biometrika* 1959 Dec;46(3/4):306. [doi: [10.2307/2333528](https://doi.org/10.2307/2333528)]
27. Sandgren N, Stoica P, Babu P. On moving average parameter estimation. *Eur Signal Process Conf* 2012:2348-2351.

28. Żebrowska M, Dzwiniel P, Waleszczyk WJ. Removal of the sinusoidal transorbital alternating current stimulation artifact from simultaneous EEG recordings: effects of simple moving average parameters. *Front Neurosci* 2020 Jul 29;14:735 [FREE Full text] [doi: [10.3389/fnins.2020.00735](https://doi.org/10.3389/fnins.2020.00735)] [Medline: [32848538](https://pubmed.ncbi.nlm.nih.gov/32848538/)]
29. van Staa T, Abenhaim L, Leufkens H. A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. *J Clin Epidemiol* 1994 Feb;47(2):183-189. [doi: [10.1016/0895-4356\(94\)90023-x](https://doi.org/10.1016/0895-4356(94)90023-x)]
30. Vlachos B, Mata-Cases M, Mundet-Tuduri X, Vallès-Callol J, Real J, Farre M, et al. Analysis of the adherence and safety of second oral glucose-lowering therapy in routine practice from the mediterranean area: a retrospective cohort study. *Front Endocrinol (Lausanne)* 2021 Jul 14;12:708372 [FREE Full text] [doi: [10.3389/fendo.2021.708372](https://doi.org/10.3389/fendo.2021.708372)] [Medline: [34335477](https://pubmed.ncbi.nlm.nih.gov/34335477/)]
31. Franklin JM, Gopalakrishnan C, Krumme AA, Singh K, Rogers JR, Kimura J, et al. The relative benefits of claims and electronic health record data for predicting medication adherence trajectory. *Am Heart J* 2018 Mar;197:153-162. [doi: [10.1016/j.ahj.2017.09.019](https://doi.org/10.1016/j.ahj.2017.09.019)] [Medline: [29447776](https://pubmed.ncbi.nlm.nih.gov/29447776/)]
32. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results: an unreported step in pharmacoepidemiology studies. *Pharmacoepidemiol Drug Saf* 2018 Jul 17;27(7):781-788 [FREE Full text] [doi: [10.1002/pds.4440](https://doi.org/10.1002/pds.4440)] [Medline: [29667263](https://pubmed.ncbi.nlm.nih.gov/29667263/)]
33. Cid-Ruzafa J, Lacy BE, Schultze A, Duong M, Lu Y, Raluy-Callado M, et al. Linacotide utilization and potential for off-label use and misuse in three European countries. *Therap Adv Gastroenterol* 2022 Jun 11;15:17562848221100946 [FREE Full text] [doi: [10.1177/17562848221100946](https://doi.org/10.1177/17562848221100946)] [Medline: [35706826](https://pubmed.ncbi.nlm.nih.gov/35706826/)]

Abbreviations

ACE: angiotensin-converting enzyme

ATC: Anatomical Therapeutic Chemical

EHR: electronic health record

IDIAPJGol: Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina

SIDIAP: Information System for Research in Primary Care

Edited by C Lovis; submitted 14.03.22; peer-reviewed by E Urtnasan, G Yu; comments to author 19.08.22; revised version received 19.09.22; accepted 13.10.22; published 15.11.22

Please cite as:

Ouchi D, Giner-Soriano M, Gómez-Lumbreras A, Vedia Urgell C, Torres F, Morros R

Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study

JMIR Med Inform 2022;10(11):e37976

URL: <https://medinform.jmir.org/2022/11/e37976>

doi: [10.2196/37976](https://doi.org/10.2196/37976)

PMID:

©Dan Ouchi, Maria Giner-Soriano, Ainhoa Gómez-Lumbreras, Cristina Vedia Urgell, Ferran Torres, Rosa Morros. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org/>), 15.11.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

4.2 Diabetes Research and Clinical Practice

El segundo artículo cubre los objetivos 2, 3 y 4 de la tesis (descrito en Tabla 6). En este estudio se modeló la exposición a tratamientos antidiabéticos, para analizar los perfiles de tratamiento a lo largo del tiempo, hasta tercera línea de tratamiento, en pacientes recién diagnosticados con T2DM.

Artículo titulado “*Longitudinal treatment patterns in patients recently diagnosed with type 2 diabetes mellitus in Catalonia*”

Publicado en *Diabetes Research and Clinical Practice*

vol. 202, no. April, p. 110777, Aug. 2023

DOI: [10.1016/j.diabres.2023.110777](https://doi.org/10.1016/j.diabres.2023.110777)

Factor de Impacto en 2023: 5.1 (Q1)



Longitudinal treatment patterns in patients recently diagnosed with type 2 diabetes mellitus in Catalonia

Dan Ouchi^{a,b,*}, Maria Giner-Soriano^{a,b}, Carles Vilaplana-Carnerero^{a,b,c,d}, Ramon Monfa^{a,b,c}, Ferran Torres^e, Rosa Morros^{a,b,c,f}

^a Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

^b Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain

^c Plataforma SCReN, UICEC IDIAPJGol, Spain

^d Departament de Medicina, Universitat de Barcelona (UB), Barcelona, Spain

^e Unitat de Bioestadística Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain

^f Institut Català de la Salut, Barcelona, Spain

ARTICLE INFO

Keywords:

Electronic health records
Primary health care
Type 2 diabetes mellitus
Drug therapy combination
Longitudinal treatment patterns
Glycemic levels

ABSTRACT

Aim: To investigate initial and subsequent treatments prescribed to newly diagnosed type 2 diabetes mellitus (T2DM) patients.

Methods: Data from SIDIAP (Information System for Research in Primary Care) including all recorded incident T2DM patients in primary care between 2015 and 2020. We used descriptive statistics and different graphical techniques to describe the most frequent longitudinal patterns.

Results: A total of 86,854 patients were included. 78.3 % of the patients began treatment with a single metformin medication and 21.7 % began with a combination therapy (CT). Metformin was the most frequent treatment as first and third-line therapy, while the CT of metformin with DPP4i or sulfonylurea was more prevalent as second-line. Most common first to third-line pattern was initial metformin for 15 months, adding a second antidiabetic in the second line, staying in CT for 6 months, and switching back to single metformin. Treatment patterns varied depending on HbA1c levels, with higher levels (>8 %) being associated with changes to CT and lower levels with switches to monotherapy or temporary discontinuation.

Conclusion: The study described in detail the different treatment patterns in incident T2DM patients in Catalonia, its adherence to the guidelines, and how the changes are associated to the HbA1c dynamics.

1. Introduction

Paths of treatment for type 2 diabetes mellitus (T2DM) have become increasingly complex with the availability of new medication classes and accompanying updates in clinical evidence and practice guidelines.

There is a wide range of oral and parenteral drugs available to effectively manage both type 1 and T2DM. These drugs are classified based on their mechanism to reduce blood glucose levels and include insulins, biguanides, sulfonylureas, meglitinides (glinides), alpha-glucosidase inhibitors (AGIs), thiazolidinediones (TZDs), incretin-dependent therapies such as glucagon-like peptide-1 receptor agonists (GLP1) and dipeptidyl peptidase-4 inhibitors (DPP4i), SGLT2 inhibitors (SGLT2i), and various combinations of these drugs. Recent studies have

shown that newer agents such as SGLT2i and GLP1a may offer clinical benefits for patients with type 2 diabetes. These advancements include a decreased likelihood of cardiovascular events for patients with cardiovascular disease (CVD) and the ability to slow down the progression of chronic kidney disease. Conversely, the evidence supporting clinical benefits for other drug classes, such as DPP4 inhibitors (DPP4i), remains limited [1]. Furthermore, it is important to note that certain medications, like sulfonylureas, may carry the potential risk of adverse events, particularly hypoglycemia in older individuals.

Optimal control of glycated hemoglobin (HbA1c) and reducing the risk of complications in T2DM is dependent on the choice of initial therapy and subsequent treatment changes [2]. As evidence in this area evolves, practice guidelines are regularly updated to evaluate the

* Corresponding author at: Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Gran Via de les Corts Catalanes 587, Barcelona 08007, Spain.

E-mail address: douchi@idiapjgol.info (D. Ouchi).

<https://doi.org/10.1016/j.diabres.2023.110777>

Received 18 April 2023; Received in revised form 30 May 2023; Accepted 8 June 2023

Available online 14 June 2023

0168-8227/© 2023 Elsevier B.V. All rights reserved.

efficacy and safety of medications, as well as the thresholds to initiate treatment and HbA1c targets [1,3–5]. In 2019 and 2021, the Catalan guidelines were updated to reflect new evidence and treatments [6]. However, the patterns of medication usage for T2DM in Catalonia have not been evaluated recently. To gain insight into the current patient profile and drug management, contemporary data on blood glucose-lowering drug utilization are required. This information has significant public health and clinical implications as it will aid in understanding the need for implementation of local policies, education, and quality improvement initiatives for people living with T2DM in Catalonia, ultimately improving their health and wellbeing.

While clinical practice patterns for patients who initiate glucose-lowering agents have been explored in other regions with a high prevalence of T2DM [7–14], there is limited research in Catalonia regarding such patterns. Some studies have cross-sectionally described the treatments used by T2DM patients [15–17], but it remains unclear how clinicians prescribe treatments longitudinally and what the most frequent patterns are. Understanding the prescribing patterns is of importance not only within the region but also for the broader scientific community. Given the global impact of T2DM, insights into clinical practice patterns can contribute to the development of evidence-based guidelines and inform healthcare providers worldwide.

Therefore, the objectives of this project were: (1) to investigate the prevalence of the initial glucose-lowering treatment, (2) to describe the most frequent sequential patterns in the treatment of T2DM patients in a real-world setting, and (3) to examine association of the longitudinal treatment patterns and HbA1c levels.

2. Methods

A longitudinal observational study conducted using electronic health records in incident T2DM patients from primary care in Catalonia (northeastern of Spain) between 2015 and 2020. Primary care in Spain is free of charge and is the main entry for public non-emergency health-related services that are delivered through primary care. Patients access to their medication in the community pharmacies through the prescriptions issued by the specialist, and most of these records are captured by the Information System for Research in Primary Care (SIDIAP) [18,19].

SIDIAP is a comprehensive electronic health record platform that collects information from 332 primary care centers under the Catalan Health Institute (Institut Català de la Salut). It includes patient data such as disease diagnoses classified using International Classification of Disease (ICD-10) [20], drug prescriptions, clinical parameters, socio-demographic information, and it is linked to the hospital discharge database. The database covers a representative population of over 7 million individuals in Catalonia, located in the northeast of Spain, bordering France to the north, in a territory of 32,000 km², and with 7,675,217 inhabitants in 2019. The information in the SIDIAP is recorded and stored in a pseudonymized manner to protect patient privacy.

2.1. Population selection and study design

All patients who were newly diagnosed with T2DM (index date) during the follow-up (2015 to 2020) with at least one glucose-lowering prescription or dispensation record 90 days prior to or following the index date. Patients were eligible for the study if they had at least 12-month follow-up. We collected data on demographic and clinical characteristics such as age, gender, smoking habits, alcohol consumption, body mass index (BMI), creatinine levels, HbA1c, estimated glomerular filtration rate (eGFR), and information on any comorbidities and other concomitant medications. Demographic and clinical data were collected at the initiation of T2DM diagnosis, the index date. For laboratory data, the closest measurement from the past year was used, and all HbA1c measured were captured to look for variations during the follow-up. Comorbidities were any health problems recorded in the year prior to

the index date, and concomitant medications were those with at least one prescription or dispensation record within the past year. Patients were followed until death, or end of data availability (transferred to another database or December 31, 2020).

2.2. Glucose-lowering medications

Glucose-lowering medications were coded according to the Anatomical Therapeutic Chemical (ATC) codes [21]: A10A (insulin preparations) and A10B (oral glucose-lowering medication). Exposure to glucose-lowering medications was defined as a dispensing record of a prescription for the medications categorized into the following groups (ATC codes): (1) metformin (A10BA); (2) sulfonylureas (A10BB); (3) dipeptidyl peptidase 4 (DPP4i) inhibitors (A10BH); (4) glucagon-like peptide-1 (GLP1) analogues (A10BJ); (5) sodium-glucose co-transporter 2 (SGLT2i) inhibitors (A10BK); (6) other hypoglycemic agents (others, including A10BX, A10BF and A10BG) and (7) insulins (A10A) (supplementary table 1).

2.3. Electronic medication records

In SIDIAP, electronic medication records were available as both prescriptions and drug pharmacy invoices and were organized in longitudinal format. Each record comprised the pseudonymized patient identifier, the ATC code, and the date. The initiation and termination dates of prescriptions were recorded by the primary care healthcare provider, while for pharmacy invoices only the month of collection was recorded. Thus, the duration of invoice records was calculated based on the number of packages dispensed, assuming each package provided coverage for 30 days, as in our region all diabetic treatments, except metformin alone, contain the tablets/capsules necessary for one month's treatment.

2.4. Modelling longitudinal treatment patterns

We used a computational approach, called the smooth algorithm [22], to model longitudinal drug patterns both single and in combination. The algorithm employs non-parametric statistical techniques similar to time-series analysis to determine the most likely therapy based on all drug prescriptions or dispensations recorded for each patient during the study period. The algorithm parameter was established to 60 days, its default. From the algorithm we obtained the treatment pattern for each patient, including the initial treatment, its duration, any sequential changes, periods in monotherapy (exposed to one drug) or combination therapy (CT, exposed to ≥ 2 or more drugs), discontinuations, and/or termination. First-line treatment was the initial treatment after T2DM diagnosis. Afterwards, we defined a switch (to second or third-line therapy) when the medication group in the patient's possession changed from one group to one or more other groups, an add-on if a new group of medication was dispensed, while the previously dispensed medication was still in possession, and a treatment temporary discontinuation when the algorithm detected a period without medication.

2.5. Compliance

Adherence was calculated as an approximation of the medication possession ratio (MPR), which was determined as the ratio of the number of days covered by the medication dispensed at the pharmacy, to the total number of days of active prescriptions for that treatment. A threshold of 80 % was used to categorize patients as either adherent or non-adherent.

2.6. Statistical and data visualization methods

Descriptive statistics and visual representations were used to explore the longitudinal treatment patterns.

Baseline characteristics were described using the classical measures. For continuous variables, the mean and standard deviation or the median and interquartile range were calculated. Categorical variables were described using relative and absolute frequencies. Multivariable logistic regression models were utilized to examine the association between therapy change and the presence of HbA1c levels $\geq 8\%$ within the three months preceding the change. The model was adjusted for the baseline parameters and adherence to previous therapy.

Different visualization techniques were used. Boxplot and violin plot [23] to represent the distribution of continuous variables such as MPR, HbA1c levels, or the time exposed to the initial treatment; dot plots to represent most common first-line therapies and its duration (the first dot is located at the average days between diagnosis and treatment initiation, while the second dot symbolizes the median time exposed to that treatment), and to illustrate the association between therapy changes and HbA1c; upset chart [24,25] for the distribution of treatments in each therapeutic line; alluvial diagrams [26] to visualize the flow between monotherapy and CT across study; and longitudinal barplots to show the duration, changes or temporary discontinuations for the longitudinal treatment patterns.

All analyses and graphics were conducted using R statistical software version 4.2 or above with a significance level of 5 %. No methods were employed to handle missing data.

3. Results

We analyzed data from 86,854 incident patients with T2DM, which represents 60.4 % of the total of 143,856 incident T2DM patients (Fig. 1a). These patients had more than one year of follow-up (Fig. 1b), at least one antidiabetic prescription, an mean age of 58.8 years (SD = 13.5), and 58.3 % were men (supplementary table 2).

3.1. First-line therapy

Baseline characteristics of the patients in accordance with their initial treatment are presented in Table 1 (five most frequent

monotherapies and CT), and in supplementary Table 2. The preferred initial therapy for T2DM was monotherapy, 78.3 %, while the remaining 21.7 % of patients initiated with CT.

Metformin was the most common treatment ($n = 58,504$, 67.4 %), and was prescribed in patients with high BMI (BMI ≥ 30 , 57.0 %) good control of HbA1c levels ($\leq 7\%$ in 48.5 %) and a higher eGFR (≥ 60 mL/min in 91.7 %). Patients who initiated therapy with DPP4i were older (mean (SD), 66.8 years (13.7)) and had poorer control of eGFR (≤ 60 mL/min in 44.5 %). Those starting with SGLT2i were more likely to have a high BMI (BMI ≥ 30 , 68.4 %), while patients starting with insulin were less obese (BMI ≥ 30 , 33.8 %) but had poorer HbA1c control ($\geq 8\%$ in 60.6 %). The most common comorbidities observed were hypertension and dyslipidemia, which had a higher prevalence among patients initiating treatment with DPP4i. In concomitant medication, the most common was antihypertensives, oral anticoagulants and nonsteroidal anti-inflammatory drugs, more frequent in patients starting with CT and less frequent in patients starting with DPP4i.

Patients starting in CT had higher HbA1c levels (HbA1c $\geq 8\%$ in 11,783, 73.7 %) than those starting monotherapy (31,394, 52.1 %, p -value < 0.001 . Supplementary table 2). The most frequent CT consisted of metformin combined with sulfonylureas (4.3 %), insulin (4.1 %), DPP4i (4.1 %), or other antidiabetic drugs (1.9 %).

Table 2 and Fig. 2 describe the treatment characteristics in terms of time to initiation of first treatment, duration, and adherence. In all treatments, the median value for the time to initiation was 0 days as $> 50\%$ of patients started at the moment of diagnosis (Overall, 58.8 %, supplementary table 2). However, CT are, in average, prescribed earlier than monotherapies (89.62 days, SD 279.1 and 190.7 days, SD 366.2, respectively; p -value < 0.001). The earliest treatment was insulin in monotherapy, 107.39 days SD 296.19, and had long exposure (> 12 months, 51.7 %). Conversely, SGLT2i was the treatment prescribed later, 294.66 days, SD 466.08, but were the patients with shorter exposure period (median 379 days, IQR 125–634). A clearer comparison is shown in Fig. 2a. The first dot in the figure corresponds to the average time interval between diagnosis and the initiation of the initial treatment. The distance between the first and second dot, connected by a

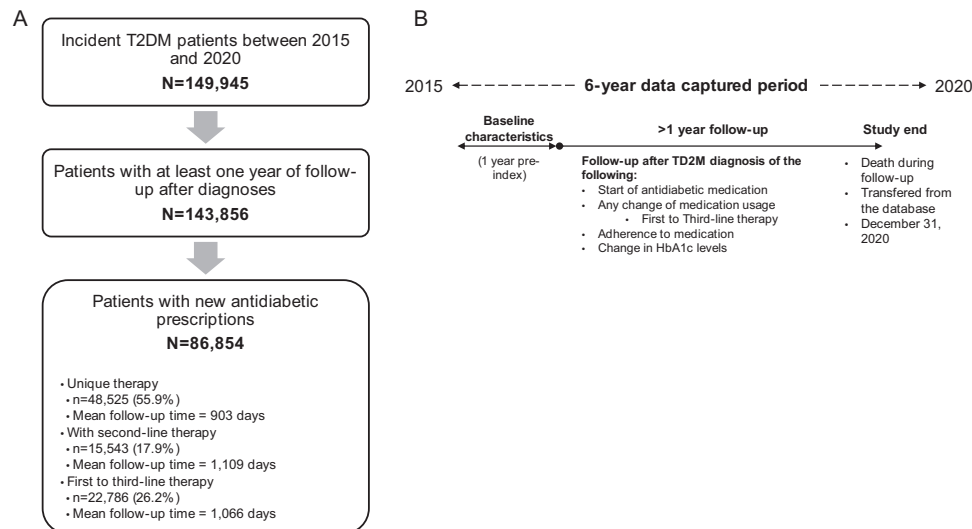


Fig 1. Flowchart of population selection and study design.

Table 1
Patient Characteristics by Initial Antidiabetic Treatment Following T2DM Diagnosis.

		Monotherapy					Combination Therapy				
		Metformin	Insulin	DPP4i	Sulfonylurea	SGLT2i	Metformin+ Sulfonylurea	Insulin+ Metformin	DPP4i+ Metformin	Metformin+ Other	Insulin+ Other
N = 86,854, %		58,504 (67.4 %)	3526 (4.1 %)	2189 (2.5 %)	1549 (1.8 %)	473 (0.5 %)	3760 (4.3 %)	3570 (4.1 %)	3537 (4.1 %)	1661 (1.9 %)	622 (0.7 %)
Clinical and demographical characteristics											
Sex (%)	<i>men</i>	33,743 (57.7)	2078 (58.9)	1136 (51.9)	729 (47.1)	264 (55.8)	2477 (65.9)	2413 (67.6)	2327 (65.8)	919 (55.3)	329 (52.9)
Age (mean (SD))		58.62 (12.91)	59.59 (16.77)	66.77 (13.74)	63.49 (14.18)	56.44 (12.27)	56.17 (12.77)	54.83 (14.24)	57.86 (13.56)	56.27 (14.32)	61.78 (15.69)
Smoker	<i>Yes</i>	29,476 (50.4)	1698 (48.2)	1005 (45.9)	660 (42.6)	241 (51.0)	1820 (48.4)	1873 (52.5)	1746 (49.4)	812 (48.9)	306 (49.2)
Alcohol habits (%)	<i>no or occasional</i>	23,577 (54.1)	1620 (69.2)	968 (62.7)	710 (64.8)	185 (58.4)	1546 (55.3)	1686 (62.3)	1388 (54.6)	769 (65.6)	324 (74.0)
26.5 %	<i>moderate</i>	18,687 (42.8)	655 (28.0)	548 (35.5)	358 (32.7)	123 (38.8)	1143 (40.9)	939 (34.7)	1060 (41.7)	370 (31.6)	109 (24.9)
missing	<i>risk</i>	1350 (3.1)	65 (2.8)	29 (1.9)	28 (2.6)	9 (2.8)	105 (3.8)	81 (3.0)	94 (3.7)	33 (2.8)	5 (1.1)
BMI, kg/m²	<i>mean (SD)</i>	31.59 (5.60)	28.28 (5.55)	30.03 (5.41)	30.53 (5.70)	33.20 (5.74)	30.86 (5.76)	30.29 (5.71)	30.54 (5.72)	33.18 (7.27)	30.88 (6.52)
23.4 %	<i>median [IQR]</i>	30.86 [27.71, 34.71]	27.69 [24.48, 31.51]	29.52 [26.40, 33.06]	29.86 [26.45, 33.65]	32.67 [29.30, 36.77]	30.08 [26.84, 34.00]	29.54 [26.26, 33.56]	29.74 [26.60, 33.51]	31.88 [28.10, 37.44]	29.94 [26.37, 34.75]
missing	<i><25</i>	4127 (9.0)	693 (29.1)	262 (16.6)	165 (14.6)	17 (5.3)	372 (13.1)	486 (17.1)	369 (14.5)	128 (10.2)	80 (17.3)
	<i>25 to 30</i>	15,518 (34.0)	885 (37.1)	590 (37.3)	411 (36.4)	84 (26.2)	1030 (36.2)	1021 (35.9)	942 (37.0)	333 (26.6)	153 (33.0)
	<i>≥30</i>	26,024 (57.0)	805 (33.8)	731 (46.2)	554 (49.0)	219 (68.4)	1442 (50.7)	1336 (47.0)	1232 (48.4)	789 (63.1)	230 (49.7)
Laboratory results (last measure in the previous year)											
HbA1c, %	<i>mean (SD)</i>	7.48 (1.47)	8.45 (2.65)	7.37 (1.38)	7.64 (1.60)	7.24 (1.36)	8.79 (2.15)	9.41 (2.83)	8.67 (2.15)	8.38 (2.31)	8.33 (2.14)
12.1 %	<i>median [IQR]</i>	7.10 [6.50, 8.00]	7.70 [6.40, 10.00]	7.10 [6.50, 7.90]	7.30 [6.60, 8.30]	6.90 [6.40, 7.60]	8.50 [7.00, 10.30]	9.20 [6.80, 11.60]	8.30 [7.00, 10.10]	7.90 [6.50, 9.80]	7.90 [6.90, 9.40]
missing	<i>≥8 %</i>	27,187 (51.5)	1557 (60.6)	977 (51.9)	764 (57.8)	176 (45.4)	2408 (74.0)	2111 (71.4)	2183 (73.1)	894 (64.3)	359 (70.7)
eGFR, mL/min	<i>mean (SD)</i>	81.29 (12.59)	71.50 (22.96)	63.07 (23.11)	76.37 (16.73)	78.66 (15.17)	82.43 (12.79)	81.58 (14.06)	81.02 (13.94)	78.71 (16.17)	62.98 (24.99)
11.5 %	<i>median [IQR]</i>	89.00 [75.10, 90.10]	82.75 [55.10, 90.10]	65.20 [42.82, 86.88]	83.80 [65.93, 90.10]	85.70 [69.80, 90.10]	90.10 [78.40, 90.10]	90.10 [77.40, 90.10]	90.10 [75.40, 90.10]	90.10 [70.50, 90.10]	66.35 [40.00, 90.10]
missing	<i><15</i>	14 (0.0)	52 (1.8)	26 (1.3)	2 (0.1)	0 (0.0)	1 (0.0)	1 (0.0)	4 (0.1)	0 (0.0)	7 (1.3)
	<i>15–29</i>	75 (0.1)	158 (5.5)	145 (7.5)	12 (0.9)	2 (0.5)	11 (0.3)	17 (0.6)	11 (0.4)	6 (0.4)	52 (10.0)
	<i>30–44</i>	868 (1.6)	300 (10.5)	356 (18.4)	78 (5.8)	20 (5.1)	49 (1.5)	86 (2.8)	70 (2.4)	80 (5.8)	102 (19.6)
	<i>45–59</i>	3469 (6.5)	296 (10.4)	332 (17.2)	152 (11.3)	36 (9.1)	196 (6.2)	204 (6.7)	214 (7.3)	130 (9.4)	73 (14.0)
	<i>≥60</i>	48,633 (91.7)	2052 (71.8)	1071 (55.5)	1106 (81.9)	337 (85.3)	2923 (91.9)	2727 (89.9)	2628 (89.8)	1168 (84.4)	286 (55.0)
Comorbidities											
	Hypertension (%)	33,146 (56.7)	1796 (50.9)	1524 (69.6)	963 (62.2)	268 (56.7)	1859 (49.4)	1626 (45.5)	1795 (50.7)	936 (56.4)	408 (65.6)
	Dyslipidemia (%)	28,323 (48.4)	1325 (37.6)	1083 (49.5)	692 (44.7)	231 (48.8)	1545 (41.1)	1325 (37.6)	1389 (38.9)	1509 (42.7)	714 (43.0)
	Acute Ischemic Stroke (%)	395 (0.7)	29 (0.8)	23 (1.1)	17 (1.1)	1 (0.2)	16 (0.4)	27 (0.8)	27 (0.8)	24 (1.4)	6 (1.0)
	Arterial hypertension (%)	1994 (3.4)	258 (7.3)	149 (6.8)	65 (4.2)	18 (3.8)	160 (4.3)	187 (5.2)	139 (3.9)	112 (6.7)	85 (13.7)
	Cerebral Vascular Accident (stroke) (%)	1780 (3.0)	227 (6.4)	120 (5.5)	62 (4.0)	15 (3.2)	116 (3.1)	149 (4.2)	116 (3.3)	76 (4.6)	53 (8.5)
	Cancer (%)	5778 (9.9)	619 (17.6)	415 (19.0)	208 (13.4)	47 (9.9)	264 (7.0)	300 (8.4)	321 (9.1)	149 (9.0)	104 (16.7)

(continued on next page)

Table 1 (continued)

	Monotherapy					Combination Therapy				
	Metformin	Insulin	DPP4i	Sulfonylurea	SGLT2i	Metformin+ Sulfonylurea	Insulin+ Metformin	DPP4i+ Metformin	Metformin+ Other	Insulin+ Other
Ischemic cardiopathy (%)	3968 (6.8)	394 (11.2)	276 (12.6)	122 (7.9)	75 (15.9)	281 (7.5)	302 (8.5)	260 (7.4)	187 (11.3)	107 (17.2)
Atrial fibrillation (%)	2977 (5.1)	377 (10.7)	319 (14.6)	140 (9.0)	33 (7.0)	137 (3.6)	172 (4.8)	211 (6.0)	95 (5.7)	71 (11.4)
Heart failure (%)	1622 (2.8)	356 (10.1)	259 (11.8)	88 (5.7)	34 (7.2)	97 (2.6)	153 (4.3)	137 (3.9)	77 (4.6)	78 (12.5)
Chronic Obstructive Pulmonary Disease (%)	3670 (6.3)	330 (9.4)	245 (11.2)	89 (5.7)	26 (5.5)	163 (4.3)	206 (5.8)	234 (6.6)	122 (7.3)	70 (11.3)
Concomitant drug										
Oral anticoagulant (%)	55,414 (94.7)	3176 (90.1)	1868 (85.3)	1419 (91.6)	434 (91.8)	3604 (95.9)	3394 (95.1)	3312 (93.6)	1568 (94.4)	555 (89.2)
Antihypertensive drug (%)	57,851 (98.9)	3384 (96.0)	2062 (94.2)	1525 (98.5)	463 (97.9)	3702 (98.5)	3522 (98.7)	3490 (98.7)	1611 (97.0)	570 (91.6)
Nonsteroidal anti-inflammatory drug (%)	53,359 (91.2)	3220 (91.3)	2007 (91.7)	1417 (91.5)	424 (89.6)	3510 (93.4)	3304 (92.5)	3292 (93.1)	1530 (92.1)	560 (90.0)
Antinuclear antibodies (%)	51,209 (87.5)	2991 (84.8)	1712 (78.2)	1334 (86.1)	420 (88.8)	3336 (88.7)	3161 (88.5)	3136 (88.7)	1401 (84.3)	476 (76.5)
Antiplatelet agent (%)	49,515 (84.6)	2742 (77.8)	1650 (75.4)	1273 (82.2)	370 (78.2)	3120 (83.0)	2842 (79.6)	2946 (83.3)	1221 (73.5)	408 (65.6)
Angiotensin II receptor blocker (%)	49,346 (84.3)	3094 (87.7)	1716 (78.4)	1286 (83.0)	364 (77.0)	3236 (86.1)	3119 (87.4)	2998 (84.8)	1328 (80.0)	485 (78.0)
Beta-blocker (%)	48,820 (83.4)	2780 (78.8)	1529 (69.8)	1247 (80.5)	335 (70.8)	3223 (85.7)	3071 (86.0)	2946 (83.3)	1324 (79.7)	443 (71.2)
Diuretic (%)	49,263 (84.2)	2593 (73.5)	1490 (68.1)	1197 (77.3)	387 (81.8)	3289 (87.5)	3043 (85.2)	3067 (86.7)	1348 (81.2)	401 (64.5)
Statin (%)	34,841 (59.6)	2355 (66.8)	1116 (51.0)	958 (61.8)	265 (56.0)	2337 (62.2)	2218 (62.1)	2135 (60.4)	897 (54.0)	303 (48.7)
Proton-pump inhibitor (%)	41,903 (71.6)	1912 (54.2)	1080 (49.3)	989 (63.8)	281 (59.4)	2872 (76.4)	2552 (71.5)	2579 (72.9)	1057 (63.6)	266 (42.8)
Angiotensin-converting enzyme inhibitor (%)	40,912 (69.9)	2738 (77.7)	1582 (72.3)	1082 (69.9)	351 (74.2)	2759 (73.4)	2610 (73.1)	2615 (73.9)	1152 (69.4)	465 (74.8)

Only the 5 most frequent in monotherapy and combination treatment are shown.

line, illustrates the duration of exposure to that specific treatment.

Adherence to treatment was found to be adequate (Fig. 2b). In overall, we found 48.6 % of patients with MPR $\geq$ 80 %. The higher MPR were observed in patients initiating with DPP4i and SGLT2i in monotherapy (median 90.1 %, IQR 79.3–100.00 and 100.00 %, IQR 84.2–100.00, respectively). Additionally, a significant correlation was found between compliance and the time exposed to the treatment, with longer exposure leading to higher MPR values (Supplementary Fig. 1).

Differences in first-line therapy between genders was studied (supplementary table 2). We observed more men in general, except for those starting with sulfonylureas (52.9 % women). Women who received SGLT2i had better control of HbA1c compared to men (HbA1c $\leq$ 7 %, in 46.4 % of women vs 37.3 % in males), and SGLT2i was prescribed earlier in women (56.0 % prescribed at diagnosis against 46.2 % in men).

3.2. Patients flow between therapeutic lines

Of all patients who started antidiabetic treatment, 48,525 (55.9 %) remained on the first-line throughout the study period (85.8 % on monotherapy and 14.2 % on CT), while 38,329 (44.2 %) patients changed their therapy. 15,543 (17.9 %) of patients required a second-line therapy, while 22,786 (26.2 %) progressed to a third-line therapy (Fig. 1a and supplementary table 3).

The Sankey diagram show the flow of patients between monotherapy and CT throughout the study (Fig. 3). Among those patients changing to a second-line therapy, the most frequent was switching from monotherapy to CT (13,473/38,329, 35.2 % of cases), or receiving two monotherapies with a temporary discontinuation in between (8,668/

38,329, 22.6 %). In contrast, patients who initiated treatment with CT typically continued with other combined therapies (values in Supplementary Table 3).

Thus, the most common approach as second-line therapy was CT, accounting for 52.3 % of cases, percentages that changed based on patient baseline characteristics (Table 3). Adherents to their first-line therapy (MPR $\geq$ 80 %) or those with uncontrolled HbA1c were more likely to receive CT in the second-line therapy (62.1 % and 59.9 %, respectively), while patients over the age of 70 received CT at a lower rate (42.9 %). The duration of the second-line therapy was, in median, 9.77 months, IQR 3.03–20, which was significantly shorter than the median duration of the first-line therapy (18.33 months, IQR 6.97–34.43, p-value < 0.001).

In the third-line therapy, the distribution between monotherapy and CT was similar (55 % and 45 %, respectively), and the duration was shorter than initial therapy (median 10.3 months, IQR 3.3–20.5). Also, 9,764 patients (25.5 %) were in temporary discontinuation between first and third-line therapies.

Supplementary Fig. 2 presents an upset chart that describes the prevalence of antidiabetic groups within each therapeutic line. Metformin was the most commonly used treatment in the first and third lines, while the combination of metformin with either DPP4i or sulfonylurea was more prevalent in the second-line. Second-line therapy was found to be more complex, five out of the top 10 most prescribed treatments were CT. In contrast, the distribution of antidiabetic groups in the third-line was more similar to that observed in the first-line.

Table 2
Characteristics of Initial Antidiabetic Treatment, Duration, and Compliance.

	Monotherapy					Combination Therapy				
	Metformin	Insulin	DPP4i	Sulfonylurea	SGLT2i	Metformin + Sulfonylurea	Insulin + Metformin	DPP4i + Metformin	Metformin + Other	Insulin + Other
N = 86,854, %	58,504 (67.4 %)	3526 (4.1 %)	2189 (2.5 %)	1549 (1.8 %)	473 (0.5 %)	3760 (4.3 %)	3570 (4.1 %)	3537 (4.1 %)	1661 (1.9 %)	622 (0.7 %)
Time from T2DM diagnosis to treatment	196.43 (369.22)	107.39 (296.19)	223.32 (397.22)	159.98 (335.69)	294.66 (466.08)	114.45 (310.50)	77.97 (260.34)	134.87 (326.66)	72.39 (257.74)	27.27 (158.09)
mean (SD)										
median	0.00 [0.00, 138.00]	0.00 [0.00, 0.00]	0.00 [0.00, 166.00]	0.00 [0.00, 6.00]	0.00 [0.00, 384.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
at diagnostic (%)	31,612 (54.0)	2233 (63.3)	1153 (52.7)	908 (58.6)	239 (50.5)	2600 (69.1)	2320 (65.0)	2333 (66.0)	1296 (78.0)	1282 (91.6)
first three months (%)	8494 (14.5)	712 (20.2)	289 (13.2)	237 (15.3)	45 (9.5)	536 (14.3)	854 (23.9)	507 (14.3)	186 (11.2)	59 (4.2)
>3 months after diagnosis (%)	18,398 (31.4)	581 (16.5)	747 (34.1)	404 (26.1)	189 (40.0)	624 (16.6)	396 (11.1)	697 (19.7)	179 (10.8)	58 (4.1)
Days exposed to first treatment	747.46 (564.05)	579.11 (583.08)	603.32 (496.97)	583.10 (523.09)	473.86 (449.84)	621.18 (532.36)	505.85 (528.06)	617.65 (529.18)	539.39 (535.72)	903.36 (718.09)
mean (SD)										
median	631.00 [303.00, 1118.00]	382.50 [100.00, 866.00]	476.00 [180.00, 910.00]	440.00 [145.00, 846.00]	379.00 [125.00, 634.00]	460.50 [181.00, 916.00]	329.50 [101.25, 709.75]	477.00 [180.00, 897.00]	379.00 [128.00, 716.00]	275.00 [94.00, 600.00]
>12 months	40,246 (68.8)	1824 (51.7)	1311 (59.9)	885 (57.1)	243 (51.4)	2248 (59.8)	1654 (46.3)	2141 (60.5)	853 (51.4)	975 (69.7)
MPR first treatment	73.65 (30.35)	54.08 (35.02)	82.52 (29.43)	70.51 (32.81)	83.44 (30.58)	45.91 (40.34)	48.41 (34.80)	17.34 (33.95)	36.96 (40.67)	30.42 (38.64)
mean (SD)										
median	86.66 [57.21, 99.66]	56.59 [23.20, 87.53]	99.05 [79.30, 100.00]	85.71 [51.27, 97.30]	100.00 [84.23, 100.00]	45.90 [0.00, 88.59]	47.04 [15.93, 80.35]	0.00 [0.00, 2.65]	15.17 [0.00, 80.41]	0.00 [0.00, 67.38]
Adherent to the first treatment (%)	33,444 (57.2)	1093 (31.0)	1625 (74.2)	861 (55.6)	366 (77.4)	1212 (32.2)	901 (25.2)	449 (12.7)	421 (25.3)	270 (19.3)
MPR ≥ 80 %										

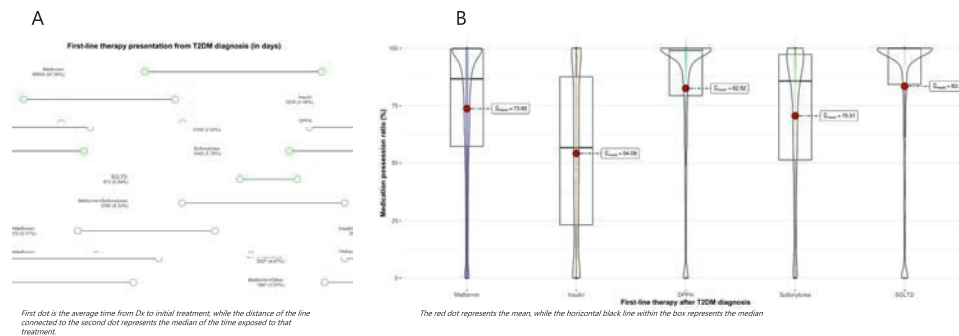


Fig 2. Duration and compliance of first-line therapy.

3.3. Longitudinal treatment patterns

We conducted a detailed analysis of the most common longitudinal patterns, incorporating the time dimension and stratifying by antidiabetic medication categories (Fig. 4).

Among the patients who did not change their therapeutic line, metformin was the most used medication with a median duration of 27 months (IQR 14–45), and was administered 6 months after diagnosis.

Next, we observed a pattern with two periods of single metformin, an initial of 6 months (IQR 3–14), with a temporary discontinuation of 4 months (IQR 2–12), and then a return to single metformin for a median of 13 months (IQR 5–25). Insulin users, on the other hand, started closer to diagnosis and had a duration, in median, of 23 months (IQR 12–42). DPP4i and combination of DPP4i with metformin were also treatments with similar durations; 22 months (IQR 12–35), and 24 months (IQR 15–39), respectively (Fig. 4A).

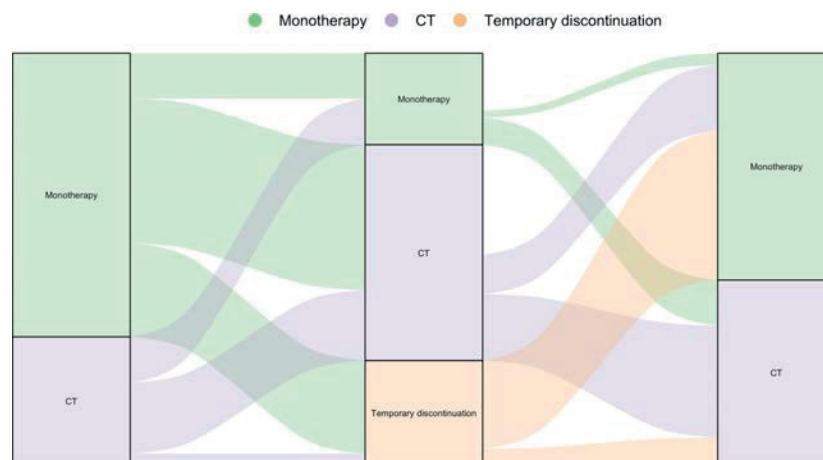


Fig 3. Sankey Diagram illustrating the flow of patients from first-line to third-line therapy.

Table 3
Distribution of Therapeutic Lines by Patient Characteristics.

	Initial therapy			Second-line therapy				Third-line therapy			
	Overall	Elderly (>70y)	HbA1c ≥ 8 %	Overall	Elderly (>70y)	Adherent* (MPR ≥ 80 %)	HbA1c ≥ 8 %	Overall	Elderly (>70y)	Adherent* (MPR ≥ 80%)	HbA1c ≥ 8 %
n (%)	86,854	17,589 (20.3)	43,177 (56.6)	38,329	6788 (17.7)	16,624 (43.4)	21,883 (57.1)	22,786	4049 (17.8)	9124 (40.0)	12,755 (56.0)
Monotherapy	68,011 (78.3)	14,060 (79.9)	31,394 (72.7)	8521 (22.2)	2116 (31.2)	2916 (17.5)	4493 (20.5)	12,528 (55.0)	2510 (62.0)	5147 (56.4)	6014 (47.2)
CT	18,843 (21.7)	3529 (20.1)	11,783 (27.3)	20,044 (52.3)	2914 (42.9)	10,328 (62.1)	13,109 (59.9)	10,258 (45.0)	1539 (38.0)	3977 (43.6)	6741 (52.8)
Temporary discontinuation	–	–	–	9764 (25.5)	1758 (25.9)	3380 (20.3)	4281 (19.6)	–	–	–	–
Duration, months median [IQR]	18.33 [6.97, 34.43]	21.10 [7.93, 39.03]	17.03 [6.43, 32.43]	9.77 [3.03, 20.00]	9.82 [3.03, 20.84]	10.33 [3.00, 20.97]	10.27 [3.27, 20.50]	10.50 [2.77, 22.90]	10.97 [3.73, 22.37]	11.20 [3.90, 22.37]	10.03 [3.47, 19.77]
Months to therapy change, median [IQR]	9.10 [3.23, 19.97]	8.78 [3.17, 20.50]	9.30 [3.33, 19.63]	5.73 [2.27, 13.50]	5.40 [2.13, 13.20]	5.27 [2.07, 13.00]	5.97 [2.37, 13.58]	–	–	–	–

* Adherent to the previous therapy.

Fig. 4B illustrates the treatment patterns for patients receiving only two lines of therapy. The predominant patterns were initiating with metformin and switching to CT. The most common pattern involved adding DPP4i to metformin, observed in 3 % of cases, with a median duration of 20 months (IQR 9–34). Those switching to DPP4i or Sulfonylurea in monotherapy had a shorter initial period of metformin, 9 months (IQR 3–23) and 6 months (IQR 6–30), respectively (Fig. 4B).

The most common first to third-line pattern was starting with single metformin, adding a second antidiabetic in the second line, and then switching back to single metformin. Of these, a frequent add-on was DPP4i after a median of 12 months (IQR 5–23) in monotherapy, and they were exposed for 6 months in median (IQR 3–12), before returning to single metformin (Fig. 4C). On the other hand, there were patterns of therapies in which patients started with a combination of metformin and another antidiabetic (DPP4i, sulfonylurea, or insulin) for about a year before removing it in the second-line therapy, leaving single metformin for 3–5 months before switching back to the initial treatment (Fig. 4C).

We had also observed patterns of therapies (not shown in Fig. 4) consisted by two consecutive add-ons, such as metformin transitioning

to a combination of metformin with DPP4i and then to a triple therapy of metformin with DPP4i and sulfonylurea ($n = 295$, 0.34 %), or metformin transitioning to metformin with sulfonylurea followed by metformin with DPP4i and sulfonylurea ($n = 287$, 0.33 %).

3.4. Association with HbA1c levels

The association between HbA1c levels three months prior to any treatment change and the subsequent switching to monotherapy, combination therapy (CT), or temporary discontinuation was examined. Fig. 5A displays a violin plot depicting the distribution of HbA1c levels in the three months preceding treatment change, categorized by the type of change (monotherapy, CT, or temporary discontinuation). Before switching to monotherapy, the median HbA1c values were 7 % (53 mmol/mol; IQR of 6.5–8 %), significantly lower than in those who changed to CT, 8.4 % (68 mmol/mol; IQR 7.3–10 %), but higher than those going to a temporary discontinuation, 6.7 % (50 mmol/mol; IQR 6.2–7.6 %). The multivariable logistic regression (forest plot in Fig. 5A) showed that patients who switched to monotherapy had a lower

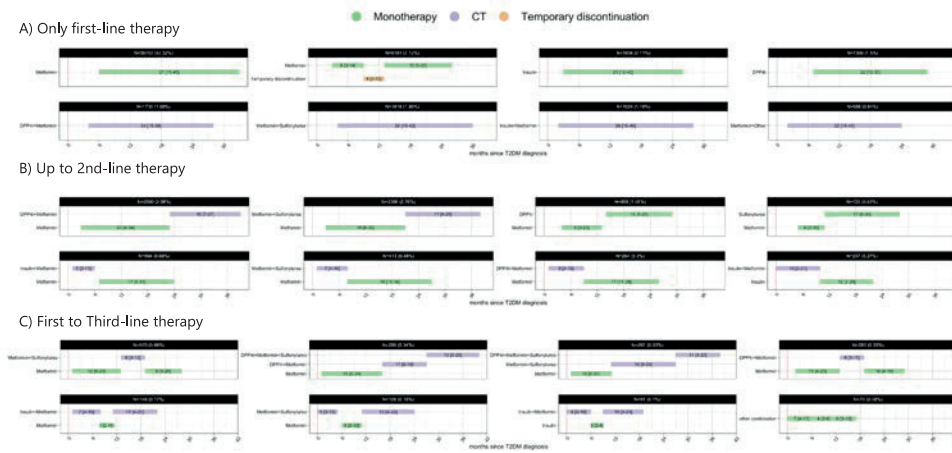


Fig 4. Longitudinal treatment pattern over time.

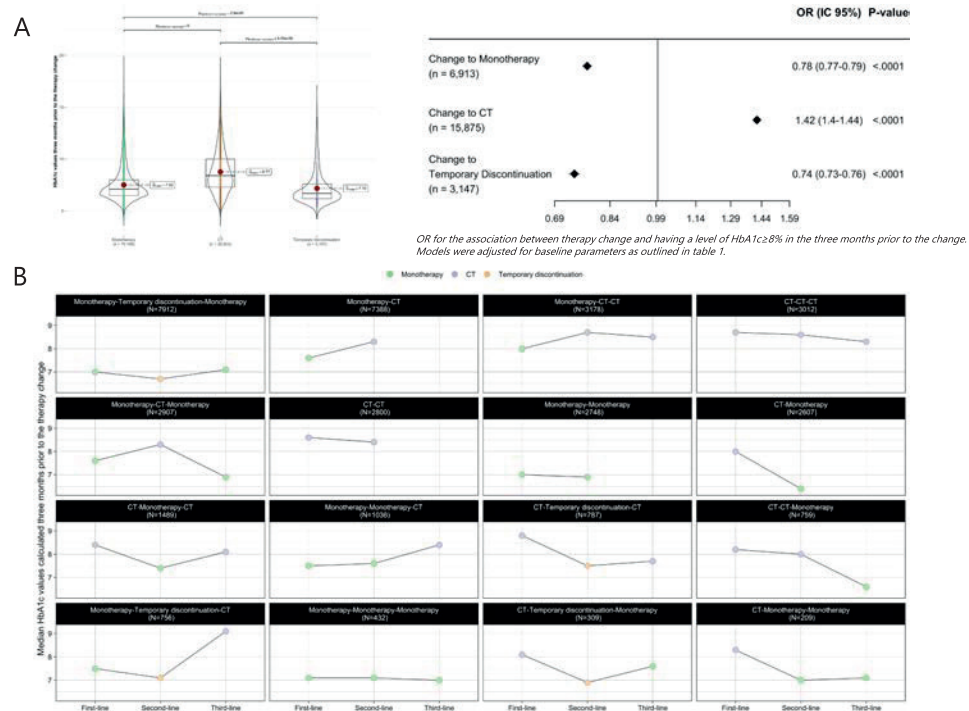


Fig 5. HbA1c levels in the three months prior to switching to a new therapeutic line.

adjusted odds-ratio (aOR) of presenting a HbA1c $\geq 8\%$ (0.78, CI 0.77–0.79), while those who changed to combination therapy (CT) had a significantly higher aOR of 1.42 (CI, 1.40–1.44) of having a HbA1c $> 8\%$ three months before the switch. Patients who changed to temporary discontinuation had a lower OR (0.74, CI 0.73–0.76) of having an uncontrolled HbA1c.

Fig. 5B also shown this relation but in a longitudinal way (values in supplementary table 3). Across patterns, an increase in HbA1c levels led to CT, while a decrease resulted in changes to monotherapy or temporary discontinuation. Patients on multiple CT treatments had consistently high HbA1c levels, while those on multiple monotherapies maintained low and stable levels. Also, each temporary discontinuation shown a previous a decrease in HbA1c levels.

4. Discussion

We have analyzed a large set of incident group of patients with T2DM from SIDIAP database to describe the first antidiabetic drug prescribed, the longitudinal patterns of T2DM treatment over time, and its association with glycemic control measured by HbA1c. We included 86,854 patients, 78.3 % of them initiated monotherapy for T2DM treatment, and metformin was the most frequently prescribed drug (67.4 %), as in other similar studies [7–14,16] and in line with current recommendations for T2DM management. [1,6,27,28].

In our study, there were 4.1 % of patients who initiated insulin as first line therapy. They were the earliest in initiating treatment after diagnosis and had the highest HbA1c of those starting monotherapy. Other studies have shown even higher percentages of patients initiating insulin as first-line treatment for T2DM. [7,10,12,29] but similar to the percentage reported in SIDIAP between 2007 and 2018 by Mata-Cases et al. [16] It is advised to initiate insulin in people who present blood glucose levels ≥ 300 mg/dL or HbA1c $> 10\%$ at the time of diagnosis, or when there are symptoms of hyperglycemia or evidence of catabolism. [30,31].

Despite being the monotherapy the recommended choice for first line treatment [6,31,32], 21.7 % of patients in our study started a CT as the first line, probably because they had poorer metabolic control than their counterparts in monotherapy ($> 73\%$ of them with HbA1c $\geq 8\%$) in line with other studies conducted in Catalonia [16]. In fact, there are data to support initial CT to attain more rapidly glycemic goals [30] and it should be considered in those presenting HbA1c levels 1.5–2 % above target. [31] Other studies have also reported CT as first line treatment, but not as frequently as in our study, for instance Carney et al. reported 6.9 % of CT with metformin as first treatment, or Nicolucci et al. described the combination of metformin and sulfonylureas in 14.6 % of newly treated patients. [9,11].

For the second line and subsequent therapies, as T2DM is mainly a progressive disease, the recommendations consist in intensification of the treatment, traditionally using a stepwise addition of medications to metformin in order to maintain the HbA1c target. [1,6 27] In our study, 44.1 % changed the first antidiabetic therapy to CT, mostly. Nevertheless, we found an unexpected frequent pattern (25.5 %) consisting in a temporary discontinuation of the antidiabetic treatment after having received a first line therapy, and a posterior return to monotherapy in most cases. We also found 5,401 (6.2 %) patients initiating CT who reduced the treatment to monotherapy (Fig. 3). This reduction of treatment does not respond to any T2DM guidelines and might be multifactorial.

As Wang [7] and Montvida et al., [10] we have also analyzed HbA1c levels during the study period to assess their association with treatment patterns (Fig. 5) and have clearly found that any change on therapy is preceded by a change in HbA1c; so that if it increased, the new treatment consisted in a treatment intensification, as recommended to maintain glycemic goals. [31,33] When HbA1c decreased, the result was a change to monotherapy or a temporary discontinuation, and this might be also recommended in some instances, such as ineffectiveness, side effects,

expenses or a change in glycemic goals. [28].

Few database studies have analysed the T2DM treatment patterns longitudinally [7,9,13], including treatment changes over time, and the association to HbA1c levels. As seen in Fig. 4 in our study, some treatment changes occurred after 6 months or multiples of 6, probably reflecting the time-frequency recommended in our setting to conduct HbA1c test in T2DM, [6] and hardly obeying to other possible causes for treatment switch such as side effects, which would have caused this switch earlier. A lack of adherence to treatment might have led to HbA1c increase and consequently, a treatment add-on, as for example 41 % of metformin-treated had suboptimal adherence (MPR $< 80\%$), thus, in the end, is the HbA1c which leads to a therapeutic modification.

Our study used the smooth algorithm [23] to model longitudinal treatment patterns, which allowed for a comprehensive analysis of treatment changes, the moment of these changes, and the degree of adherence to the guidelines. We were also able to examine the associations between these patterns and HbA1c levels or other factors. With our approach, we were able to provide valuable insights from electronic health records and medication records offering a detailed understanding of treatment scenarios that are not typically observed in the literature, highlighting its usefulness in clinical research.

This investigation has several limitations that need to be acknowledged. The study was limited to a single healthcare system, which may not be representative of other healthcare settings or populations, and it has a relatively short follow-up period which may limit the ability to draw conclusions about the different treatment patterns. Also, the proximity of the diagnostic date to the start of treatment suggests that the exact moment of disease onset may be imprecise or not well recorded. To address this issue, we limited the patient selection to those with prescription record 90 days prior to or following the diagnoses date, but this approach has limitations and did not fully address the issue. Moreover, we do not have information on whether the changes in treatment were made by the same or different specialists, which may have an impact on treatment patterns. Additionally, we lack data on the reason for treatment changes due to adverse events, and we cannot explore the association between longitudinal treatment patterns and mortality or hospitalizations. Furthermore, it is important to note that the variation in follow-up periods among patients in our study could potentially influence the observed treatment patterns. To minimize this effect, we applied a minimum follow-up criterion of at least one year for inclusion in our study, but in our population we found that longer follow-up periods are associated with an increased likelihood of receiving a third-line therapy. Therefore, the findings of this study need to be interpreted with caution and further research is needed to address these limitations.

In summary, we provide a detailed analysis of the antidiabetic treatment in incident T2DM patients, capturing the evolution from initial prescription to subsequent changes. Our longitudinal approach allows for a comprehensive investigation of treatment patterns, including when and what changes occurred and their relation to HbA1c dynamics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

To the researchers who participated in this study.

Author contributions

All authors contributed to the study design. R.Mon., C.V., M.G.S., R. Mor. and D.O. conceived the study. M.G.S., R.Mon., C.V. and D.O.

collected the data. D.O. performed the statistical analyses. D.O., C.V., R. Mon., M.G.S. and R.Mor. were involved in the interpretation of the data. D.O. and M.G.S. wrote the manuscript and D.O. designed the tables and figures. All authors critically revised the report and approved the final version to be submitted for publication. The corresponding author confirms that he had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Ethics approval

The study was approved by the Primary Care Research and Ethics Committee of the Research Institute IDIAP Jordi Gol, Barcelona, Spain (reference number, 21/026-P, 24th February 2021).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.diabres.2023.110777>.

References

- Care D, Suppl SS. 9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2022. *Diabetes Care* 2022;45(January):S125–43. <https://doi.org/10.2337/dc22-S009>.
- Mata-Cases M, Franch-Nadal J, Real J, Mauricio D. “Glycaemic control and antidiabetic treatment trends in primary care centres in patients with type 2 diabetes mellitus during 2007–2013 in Catalonia: A population-based study.” *BMJ Open*, vol. 6, no. 10, 2016, doi: 10.1136/bmjopen-2016-012463.
- Buse JB, et al., “2019 update to: Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD).” *Diabetes Care*, vol. 43, no. 2, 2020, doi: 10.2337/dci19-0066.
- Cosentino F, et al., “2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD.” *Eur Heart J*, vol. 41, no. 2, 2020, doi: 10.1093/eurheartj/ehz486.
- Care D, Suppl SS. “6. Glycemic Targets : Standards of Medical Care in Diabetes — 2022,” vol. 45, no. January, pp. 83–96, 2022.
- Fau E, et al. Pautes per al tractament farmacològic de la diabetis mellitus tipus 2 Versió del 21 d'abril de 2021. Scientia 2021;3.
- Wang C, Gao Y, Zhu L, Huang M, Wu Y, Xuan J. Treatment Patterns in Patients With Newly Diagnosed Type 2 Diabetes in China: A Retrospective, Longitudinal Database Study. *Clin Ther* 2019;41(8):1440–52. <https://doi.org/10.1016/j.clinthera.2019.05.003>.
- Sharma M, Nazareth I, Petersen I. “Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: A retrospective cohort study.” *BMJ Open*, vol. 6, no. 1, 2016, doi: 10.1136/bmjopen-2015-010210.
- Carney G, et al. Treatment pattern trends of medications for type 2 diabetes in British Columbia, Canada. *BMJ Open Diabetes Res Care* 2022;10(6):1–12. <https://doi.org/10.1136/bmjopen-2022-002995>.
- Montvida O, Shaw J, Atherton JJ, Stringer F, Paul SK. Long-term trends in antidiabetes drug usage in the U.S.: Real-world evidence in patients newly diagnosed with type 2 diabetes. *Diabetes Care* 2018;41(1):69–78. <https://doi.org/10.2337/dci17-1414>.
- Nicolucci A, et al. “Treatment patterns and associated factors in 14 668 people with type 2 diabetes initiating a second-line therapy: Results from the global DISCOVER study programme.” *Diabetes Obes Metab* 2019;21(11):2474–85. <https://doi.org/10.1111/dom.13830>.
- Van Den Heuvel JM, Farzan N, Van Hoek M, Maitland-Van Der Zee AH, Ahmadizar F. Mining treatment patterns of glucose-lowering medications for type 2 diabetes in the Netherlands. *BMJ Open Diabetes Res Care* 2020;8(1):1–7. <https://doi.org/10.1136/bmjopen-2019-000767>.
- Lee KA, Jin HY, Kim YJ, Im YJ, Kim EY, Park TS. Treatment Patterns of Type 2 Diabetes Assessed Using a Common Data Model Based on Electronic Health Records of 2000–2019. *J Korean Med Sci* 2021;36(36):1–12. <https://doi.org/10.3346/jkms.2021.36.e230>.
- Tajima, A et al., “Treatment patterns and satisfaction in patients with type 2 diabetes newly initiating oral monotherapy with antidiabetic drugs in Japan: results from the prospective Real-world Observational Study on Patient Outcomes in Diabetes (RESPOND).” *BMJ open diabetes Res. care*, vol. 10, no. 6, 2022, doi: 10.1136/bmjopen-2022-003032.
- J. Marco Franco, J. Girbes Borrás, and R. Casañ Fernández, “Guía para la prescripción y visado de antidiabéticos,” pp. 4–9, 2015.
- Mata-Cases M, et al. Trends in the Degree of Control and Treatment of Cardiovascular Risk Factors in People With Type 2 Diabetes in a Primary Care Setting in Catalonia During 2007–2018. *Front Endocrinol (Lausanne)* 2022;12 (January):1–11. <https://doi.org/10.3389/fendo.2021.810757>.
- Gómez García MC, Franch-Nadal J, Millaruelo Trillo JM, Cos-Claramunt FX, Avila Lachica L, Buil Cosiales P. “Control glucémico y de los factores de riesgo cardiovascular en los pacientes con diabetes tipo 2 con enfermedad cardiovascular en España, y su patrón de tratamiento, en función del género: Estudio CODICE.” *Med. Fam. Semer.*, vol. 46, no. 2, 2020, doi: 10.1016/j.semerg.2019.05.005.
- Del Mar García-Gil M, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAPI). *Inform Prim Care* 2012. <https://doi.org/10.14236/jhl.v19i3.806>.
- Recalde M, et al. Data Resource Profile: The Information System for Research in Primary Care (SIDIAPI). *Int J Epidemiol* 2022;51(6):e324–36. <https://doi.org/10.1093/ije/dyaa068>.
- WHO. “WHO ICD-10 Version,” *WHO - ICD-10*, 2019.
- Anon, “WHO ATC/DDD Index 2020,” *WHO Collaborating Centre for Drugs Statistics Methodology*, 2020.
- Ouchi D, Giner-Soriano M, Gómez-Lumbreras A, Vedia Urgell C, Torres F, Morros R. “Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study,” *JMIR Med Inform*, vol. 10, no. 11, p. e37976, Nov. 2022, doi: 10.2196/37976.
- Hintze JL, Nelson RD. “Violin Plots: A Box Plot-Density Trace Synergism Statistical Computing and Graphics Violin Plots: A Box Plot-Density Trace Synergism,” *Source Am. Stat.*, vol. 52, no. 2, pp. 181–184, 1998, [Online]. Available: <http://www.jstor.org/stable/2685478%5Cnhttp://www.jstor.org/%5Cnhttp://www.jstor.org/action/showPublisher?publisherCode=astata.%5Cnhttp://www.jstor.org/>
- Lex A, Gehlenborg N, Strobel H, Vuilleumot R, Pfister H. UpSet: Visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 2014;20(12):1983–92. <https://doi.org/10.1109/TVCG.2014.2346248>.
- Conway JR, Lex A, Gehlenborg N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;33(18):2938–40. <https://doi.org/10.1093/bioinformatics/btx364>.
- Rosvall M, Bergstrom CT. “Mapping change in large networks,” *PLoS One*, vol. 5, no. 1, 2010, doi: 10.1371/journal.pone.0008694.
- Davies MJ, et al. Management of hyperglycaemia in type 2 diabetes, 2022. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia* 2022;65(12):1925–66. <https://doi.org/10.1007/s00125-022-05787-2>.
- Aleman JJ, et al. Guía de diabetes tipo 2 para clínicos. Recomendaciones de la redGDPS 2018.
- Khunti K, et al. “Patterns of glycaemic control in patients with type 2 diabetes mellitus initiating second-line therapy after metformin monotherapy: Retrospective data for 10 256 individuals from the United Kingdom and Germany.” *Diabetes Obes Metab* 2018;20(2):389–99. <https://doi.org/10.1111/dom.13083>.
- Phung OJ, Sobieraj DM, Engel SS, Rajpathak SN. “Early combination therapy for the treatment of type 2 diabetes mellitus: Systematic review and meta-analysis”. *Diabetes Obes Metab* 2014;16(5):410–7. <https://doi.org/10.1111/dom.12233>.
- Elsayed NA, et al., “9. Pharmacologic Approaches to Glycemic Treatment: Standards of Care in Diabetes—2023.” *Diabetes Care*, vol. 46, no. suppl. pp. S140–S157, 2023, doi: 10.2337/dc23-S009.
- Shin J-I. Second-line Glucose-Lowering Therapy in Type 2 Diabetes Mellitus. *Curr Diab Rep Aug.* 2019;19(8):54. <https://doi.org/10.1007/s11892-019-1171-0>.
- Of S, Carediabetes M. “American Diabetes Association. 8. Pharmacologic approaches to glycemic treatment: Standards of Medical Care in Diabetes—2018,” *Diabetes Care*, vol. 41, no. Supplement 1, 2018.

4.3 Drugs – Real World Outcomes

El tercer y último artículo de la tesis cubre los objetivos 2, 3 y 5 (descrito en Tabla 6). En este estudio se buscaron pacientes con diabetes mellitus tipo 2 (T2DM) bajo tratamiento inicial con Metformina y enfermedad no controlada, medida a partir de la hemoglobina glicada (HbA1c). El algoritmo fue necesario para determinar la exposición a antidiabéticos, buscar el momento de adición del tratamiento de segunda línea, hacer el seguimiento longitudinal y calcular el cumplimiento.

Artículo titulado *“Impact of Second-Line Combination Treatment for Type 2 Diabetes Mellitus on Disease Control: A Population-Based Cohort”*

Publicado en *Drugs - Real World Outcomes*

Aceptado: Abril 2023; Publicado: Mayo 2023

DOI: <https://doi.org/10.1007/s40801-023-00374-2>

Factor de Impacto en 2023: 2.0 (Q2)



Impact of Second-Line Combination Treatment for Type 2 Diabetes Mellitus on Disease Control: A Population-Based Cohort Study

Dan Ouchi^{1,2,3} · Carles Vilaplana-Carnerero^{1,2,3} · Ramon Monfà^{1,2,3} · Maria Giner-Soriano^{1,2,3} · Ana Garcia-Sangenís^{1,2,3} · Ferran Torres⁴ · Rosa Morros^{1,2,3,5}

Accepted: 28 April 2023
© The Author(s) 2023

Abstract

Background Type 2 diabetes mellitus is a chronic disease affecting millions of people worldwide. Achieving and maintaining glycemic control is essential to prevent or delay complications and different strategies are available as second-line treatment options for patients with type 2 diabetes who do not achieve glycemic control with metformin monotherapy.

Objective The aim of this work is to describe the impact of initiating a combination treatment to reduce glycated hemoglobin in patients with type 2 diabetes with insufficient glycemic control.

Methods We included patients with a type 2 diabetes diagnosis between 2015 and 2020 at the Information System for Research in Primary Care (SIDIAP) database in Catalonia, Spain. The primary outcome was the time to glycated hemoglobin control ($\leq 7\%$) during the first 720 days, expressed as the restricted mean survival time. Adjusted differences of the restricted mean survival time were compared to analyze the performance of each treatment versus the combination with a sulfonylurea. Adherence was calculated as the medication possession ratio using an algorithm to model treatment exposure.

Results A total of 28,425 patients were analyzed. The most frequent combinations were those with sulfonylureas and dipeptidyl peptidase-4 inhibitors. All treatments reduced glycated hemoglobin and the restricted mean survival time for the sulfonylurea treatment was 455 (451–459) days although combinations with glucagon-like peptide-1 and insulin reached glycemic control earlier, -126 days (-152 to -100 , $p < 0.001$) and -69 days (-88 to -50 , $p < 0.001$), respectively. Adherence was high in all groups apart from the insulin combination and had a significant effect in reducing glycated hemoglobin except in sodium-glucose cotransporter type 2 inhibitors and insulin. Glucagon-like peptide-1 and sodium-glucose cotransporter type 2 inhibitors showed significant reductions in weight.

Conclusions Patients achieved the glycated hemoglobin goal with second-line treatments. Glucagon-like peptide-1 and insulin combinations achieved the goal earlier than sulfonylurea combinations. Adherence significantly reduced the time to glycated hemoglobin control except for the combination with sodium-glucose cotransporter type 2 inhibitors.

✉ Dan Ouchi
douchi@idiapjgol.info

¹ Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Gran Via de les Corts Catalanes 587, 08007 Barcelona, Spain

² Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain

³ Plataforma SCReN, UICEC IDIAPJGol, Barcelona, Spain

⁴ Unitat de Bioestadística Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain

⁵ Institut Català de la Salut, Barcelona, Spain

Published online: 09 May 2023

△ Adis

Key Points

Second-line combination treatments of metformin with glucagon-like peptide-1 and insulin achieved earlier glycated hemoglobin control than metformin and sulfonylurea combinations.

Adherence to medication significantly reduced the time to glycated hemoglobin control for most treatments, except for the combination with sodium-glucose cotransporter type 2 inhibitors.

Metformin with glucagon-like peptide-1 or with sodium-glucose cotransporter type 2 inhibitors showed significant reductions in weight.

1 Introduction

Type 2 diabetes mellitus (DM2) is a disease that affects many people in our society [1]. Although it can initially be controlled with a good diet and healthy habits, most patients will require treatment for glycemic control [2, 3].

Throughout their lives, patients diagnosed with DM2 may receive different treatments to treat their condition and control glycemia [4]. In general, it is recommended to maintain glycated hemoglobin (HbA1c) below 7%, individualizing this target value according to the patient's needs and characteristics, such as comorbidities or age [5–8]. Initially, if a patient shows a satisfactory response to treatment, HbA1c levels tend to reduce, especially during the first months of treatment, and they are unlikely to return to baseline [9].

Type 2 diabetes can be managed with a range of pharmacological treatments, with metformin and lifestyle modifications being the recommended first-line therapy, although initiation with other alternatives is considered in some situations, such as other antidiabetic drugs in combination therapy [10] if patients do not achieve adequate HbA1c control with metformin monotherapy [11, 12] or in patients with an established/high risk of atherosclerotic cardiovascular disease, heart failure, and/or chronic kidney disease, when the treatment regimen should include agents that reduce the cardiorenal risk [10]. These combination therapies include sulfonylureas, dipeptidyl peptidase-4 inhibitors (DPP4i), sodium-glucose cotransporter type 2 inhibitors (SGLT2i), glucagon-like peptide-1 (GLP1) receptor antagonists, other hypoglycemic agents (others), and insulins [13–15].

A recently published meta-analysis, including 36,746 adults in 68 randomized clinical trials, concluded that all classes of oral antidiabetic drugs result in similar

reductions in HbA1c levels when added to metformin, although SGLT2i showed an additional benefit on body weight reduction [16]. Despite the availability of many classes of medication in the DM2 field, medication adherence is poor [17]. When the target is not achieved with single metformin therapy, the second line should add a medication class that considers adherence and persistence to achieve good glycemic control for the patient in the long term. In this study, we aimed to measure and compare the impact of the different combination treatments for DM2 on the reduction in HbA1c, patient adherence, and patient weight after their addition.

2 Methods

2.1 Study Design and Population

We conducted a population-based cohort study including all adult patients with an active DM2 diagnosis and metformin prescription between 2015 and 2020 in Catalonia, Spain. The main inclusion criteria were: initial treatment of metformin monotherapy followed by a combination therapy; at least two measures of HbA1c; HbA1c measure > 7% 90 days before the second-line addition; and one or more HbA1c values at least 90 days after the start of the combination therapy. All International Classification of Diseases, Tenth Revision [18] and Anatomical Therapeutic Chemical classification system [19] codes included in the study are shown in Table 1 of the Electronic Supplementary Material (ESM).

2.2 Data Source

The study data source is the Information System for Research in Primary Care (SIDIAP) [20], which captures clinical information of approximately 5.8 million people from Catalonia, Spain (around 80% of the Catalan population). This information is pseudonymized, originating from different data sources: (1) ECAP (electronic health records in primary healthcare of the Catalan Health Institute); including sociodemographic characteristics, residents in nursing homes/long-term facilities, comorbidities registered as International Classification of Diseases, Tenth Revision codes, specialist referrals, clinical parameters, unhealthy habits, sickness leave, date of death, laboratory test data, and drug prescriptions issued in primary healthcare, registered in the Anatomical Therapeutic Chemical classification system; (2) pharmacy claims data corresponding to the primary healthcare drug prescriptions; and (3) the database of diagnoses at hospital discharge [21, 22].

2.3 Outcome Measures

The primary outcome is the time (in days) to HbA1c normalization ($\leq 7\%$) after initiating the hypoglycemic combined treatment. Different secondary outcomes were considered: (1) to assess the impact of combination therapy adherence on HbA1c levels; (2) impact on weight after initiation of the combination therapy; (3) distribution of combination therapy over time; (4) duration of the single metformin treatment (number of days between the first therapy and the combination therapy); and (5) duration of the combination therapy (number of days between the first and last prescription). The follow-up time for all the outcomes started at the day of the hypoglycemic combined treatment initiation and ended at one of the following censoring events: end of study period (720 days) or disenrollment from the database or death.

2.4 Modeling Drug Exposure to Combination Treatment

We have used the records from electronic prescriptions and dispensations in pharmacies for drug exposure. Prescriptions are recorded with the exact date but invoice data from the pharmacy are recorded by month-year and not linked to the prescription. To assess this linkage, we used the smooth algorithm, developed in our group, which automatically determines at a patient level the most probable treatment (monotherapy or combination) throughout the study period [23]. The smooth algorithm was used to model the drug exposure in both prescriptions and dispensations, to obtain the exact moment of addition of combination therapy, the period exposed to that treatment, and the compliance or adherence.

The adherence was calculated as an approximation of the medication possession ratio (MPR). It was calculated as the ratio between the total days covered by the packages dispensed at the pharmacy, the numerator, and the total days of active prescriptions for that treatment, the denominator. We used a threshold of 80% to classify patients into adherents and non-adherents.

2.5 Statistical Analysis

Results are presented as percentages, means and standard deviations, or median and interquartile range (IQR). In addition, a bivariate baseline analysis was performed. Comparisons between groups in baseline variables were performed by the analysis of variance test for continuous variables and the χ^2 test for categorical variables (exact Fisher test with observed frequencies < 5). In the primary outcome, the time in days was calculated until the first time the HbA1c was

below the threshold ($\leq 7\%$). Those patients who did not achieve HbA1c control during the follow-up period were censored at the date of the last HbA1c sample.

To assess the impact of each combination treatment on HbA1c, a survival analysis was carried out using the Kaplan–Meier and restricted mean survival time (RMST) methods [24–26]. Comparison between the survival curves was undertaken using the log-rank test. Regression models were used to estimate the RMST and its standard error using the delta method [27] and we compared the difference of each RMST with the combination of metformin with sulfonylurea, our reference treatment. The RMST was adjusted by the patient's characteristics using relevant covariates and truncated to 720 days (the follow-up time). Differences between RMSTs were calculated using the identity link and reported with 95% confidence intervals (CIs) along with their p values. For the weight reduction analyses, we fitted a logistic regression model to estimate the odds ratio (OR) of losing more than 5% of weight after the combination treatment versus the same reference group. The ORs were adjusted by the same covariates, and the 95% CI and p values were calculated using classic methods. All analyses were performed with the R statistical package version 4.2 [28, 29] or above under a significant level of 0.05.

3 Results

3.1 Study Participants

From the 43,068 patients starting with a metformin treatment, a total of 28,425 individuals had an addition of a combination therapy during the 720 days of follow-up, and at least two measures of HbA1c: one HbA1c $\geq 7\%$ before the addition and at least another HbA1c result after 90 days (Fig. 1). The demographic and clinical characteristics of the population at baseline show differences between the study groups (Table 1). Overall baseline HbA1c was 8.63% [70.78 mmol/mol] (standard deviation [SD] = 1.34), but higher in patients with a combination with insulin, 10.83% [94.86 mmol/mol] (SD = 2.09). The mean age was 63.63 years (SD = 11.38), a total of 38.1% were women, and the average weight of the population was 83.64 kg (SD = 16.64). Patients with a combination therapy of GLP1 were more frequently observed in women, 52.1%, a higher weight, 107.26 kg (SD = 20.52), and were diagnosed with DM2 3.33 (SD = 3.99) years ago. Clinical characteristics were homogeneous between groups, the mean glomerular filtration rate (as CKDEPI) was 81.12 (SD = 12.92) mL/min/1.73 m², and the main comorbidities in the population were hypertension, 19,866 (69.9%), obesity, 18,311 (64.4%), and dyslipidemia, 17,814 (62.7%).

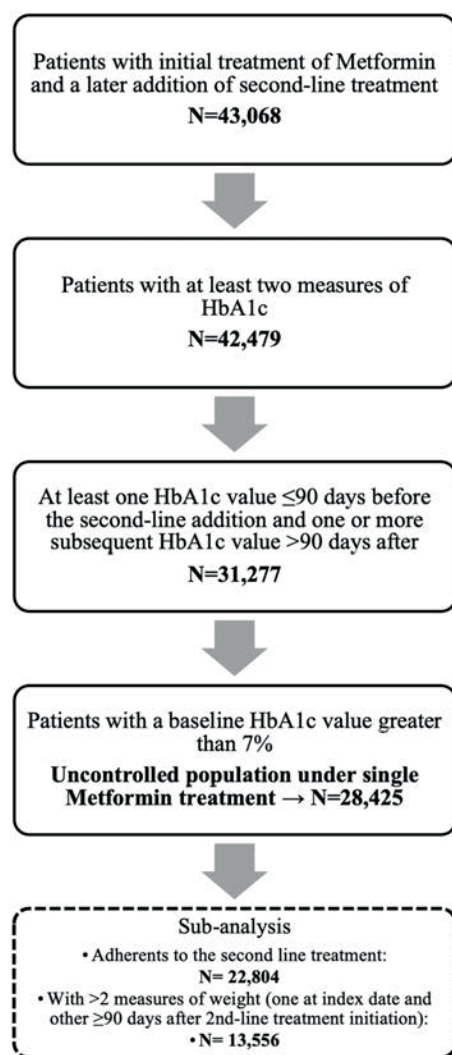


Fig. 1 Study population flowchart: participant selection and exclusion. *HbA1c* glycated hemoglobin

The most prescribed combination therapy were sulfonylureas, 13,216 (46.5%) and DPP4i, 12,000 (42.2%). The rest of the therapies were observed in 17.3% of the patients, with SGLT2i the least prescribed with only 837 (2.9%) patients. Distribution was constant over time, although

the sulfonylurea prescriptions as a combination treatment showed a negative trend, with DPP4i becoming the most prescribed from 2017 (Fig. 1 of the ESM).

3.2 Primary Outcome

On average, all treatments had a positive impact on HbA1c with a mean reduction of 1.82% (SD = 1.44, Table 2). Independently of the baseline level, we observed an exponential decay of HbA1c during the first 3–6 months after the addition of second-line treatment. Then, the HbA1c reaches its minimum before entering a slow recovery phase but without returning to baseline values (Fig. 2 of the ESM).

More than a half of the patients, 65.4%, reached the HbA1c goal during the follow-up. In the group treated with GLP1, the percentage of controlled patients was higher, 236 (81.4%), followed by insulins, 61.7%, sulfonylureas, 65.8%, DPP4i, 65.4%, and other hypoglycemic agents, 63.2%.

The efficacy on controlling HbA1c was measured using the RMST. The combination that reached the HbA1c target earlier was the GLP1 with a RMST of 330 days (95% CI 304–357), whereas intensification with other hypoglycemic agents needed more time to reach the HbA1c target, 483 days (95% CI 469–498). In patients with a combination treatment of sulfonylurea (our reference group), the RMST was 455 days (95% CI 451–459), 447 days (95% CI 442–452) in DPP4i, 452 days (95% CI 435–469) in combinations with insulins, and 457 (95% CI 440–457) in SGLT2i (Table 2).

At 720 days, the adjusted difference of RMST to the combination treatment with sulfonylureas was statistically different in all treatments except for SGLT2i ($p = 0.549$). Patients receiving GLP1 and insulins reached the HbA1c goal earlier, -126.26 (95% CI -152.49 to -100.03) and -69.18 (95% CI -88.07 to -50.30) days, respectively. While those receiving a combination treatment with other hypoglycemic agents and DPP4i needed more days to reduce the HbA1c levels, 39.01 (95% CI 24.08 – 53.95) and 14.81 (95% CI 8.61 – 21.01) days (Fig. 2).

3.3 Effect of Adherence

The MPR and frequency of adherent patients to each combination treatment were also calculated (Table 2). The mean MPR for the entire cohort was 87.76% (SD = 23.6), with a total of 80.2% adherents. The combination treatments with more adherent patients were DPP4i, 10,540 (87.8%) and sulfonylurea, 10,143 (76.7%), whereas patients prescribed with a combination with insulin were less adherent to the treatment, 465 (48.8%).

The effect of adherence on the primary outcome is shown in Table 3. At 2 years, the adjusted differences in RMST between adherent and non-adherent patients was significantly different ($p < 0.001$). Overall, adherent patients

Table 1 Demographic and clinical characteristics of the study population according to its second-line treatment

	Overall	Metformin + sulfonylurea	Metformin + GLP1	Metformin + DPP4i	Metformin + insulin	Metformin + SGLT2i	Metformin + others
<i>N</i>	28,425	13,216	290	12,000	952	837	1130
HbA1c at index date [mean (SD)]	8.6 (1.34)	8.6 (1.28)	8.6 (1.31)	8.5 (1.18)	10.8 (2.09)	8.5 (1.28)	8.6 (1.21)
HbA1c (mmol/mol) at index date [mean (SD)]	70.78 (14.70)	70.86 (14.05)	70.08 (14.34)	68.93 (12.98)	94.86 (22.89)	69.83 (14.01)	70.15 (13.27)
Years since DM2 diagnosis [mean (SD)]	5.88 (4.64)	5.77 (4.53)	3.33 (3.99)	6.14 (4.70)	5.01 (4.72)	5.10 (4.35)	6.33 (5.12)
Women (%)	10,831 (38.1)	4956 (37.5)	151 (52.1)	4588 (38.2)	359 (37.7)	326 (38.9)	451 (39.9)
Age, years [mean (SD)]	63.63 (11.38)	62.64 (11.10)	54.92 (10.87)	64.98 (11.22)	62.81 (13.47)	60.10 (10.68)	66.27 (12.03)
Smoking habit (%)							
Non-smoker	12,195 (42.9)	5579 (42.2)	99 (34.1)	5310 (44.2)	349 (36.7)	330 (39.4)	528 (46.7)
Ex-smoker	4997 (17.6)	2384 (18.0)	61 (21.0)	1992 (16.6)	203 (21.3)	178 (21.3)	179 (15.8)
Current smoker	11,233 (39.5)	5253 (39.7)	130 (44.8)	4698 (39.1)	400 (42.0)	329 (39.3)	423 (37.4)
Weight [mean (SD)]	83.64 (16.64)	84.13 (16.38)	107.26 (20.52)	82.34 (16.01)	81.09 (18.23)	90.53 (18.43)	82.70 (16.28)
BMI [mean, kg (SD)], missing 52.4%	31.39 (5.30)	31.44 (5.17)	39.81 (6.45)	31.02 (5.07)	30.42 (5.98)	34.54 (5.96)	30.95 (5.05)
BMI ≥ 30 (%)	7626 (26.8)	3705 (28.0)	123 (42.4)	2949 (24.6)	240 (25.2)	315 (37.6)	294 (26.0)
CKDEPI [mean, mL/min/1.73 m ² (SD)]	81.12 (12.92)	82.12 (12.02)	83.87 (12.22)	80.38 (13.25)	79.39 (15.07)	83.24 (10.92)	76.56 (16.71)
MDRD [mean, mL/min/1.73 m ² (SD)], missing 0.8%	59.19 (3.52)	59.40 (2.98)	59.06 (4.18)	59.11 (3.62)	58.51 (5.03)	59.58 (2.57)	57.79 (5.86)
Comorbidities, <i>N</i> (%)							
Ischemic stroke	443 (1.6)	212 (1.6)	1 (0.3)	176 (1.5)	17 (1.8)	9 (1.1)	28 (2.5)
Atherosclerosis	1994 (7.0)	909 (6.9)	12 (4.1)	836 (7.0)	97 (10.2)	51 (6.1)	89 (7.9)
Cerebrovascular disease	1644 (5.8)	750 (5.7)	17 (5.9)	698 (5.8)	65 (6.8)	36 (4.3)	78 (6.9)
Cancer	4351 (15.3)	1946 (14.7)	33 (11.4)	1907 (15.9)	165 (17.3)	104 (12.4)	196 (17.3)
Ischemic heart disease	3394 (11.9)	1547 (11.7)	25 (8.6)	1388 (11.6)	137 (14.4)	143 (17.1)	154 (13.6)
Dyslipidemia	17,814 (62.7)	8237 (62.3)	171 (59.0)	7617 (63.5)	546 (57.4)	520 (62.1)	723 (64.0)
Atrial fibrillation	2191 (7.7)	949 (7.2)	23 (7.9)	947 (7.9)	97 (10.2)	51 (6.1)	124 (11.0)
Hypertension	19,866 (69.9)	9219 (69.8)	208 (71.7)	8365 (69.7)	642 (67.4)	572 (68.3)	860 (76.1)
Heart failure	1465 (5.2)	615 (4.7)	17 (5.9)	623 (5.2)	89 (9.3)	44 (5.3)	77 (6.8)
Chronic respiratory diseases	3076 (10.8)	1234 (9.3)	21 (7.2)	1419 (11.8)	117 (12.3)	46 (5.5)	239 (21.2)
COPD	2464 (8.7)	1114 (8.4)	22 (7.6)	1020 (8.5)	123 (12.9)	77 (9.2)	108 (9.6)
Obesity	18,311 (64.4)	8753 (66.2)	283 (97.6)	7339 (61.2)	557 (58.5)	658 (78.6)	721 (63.8)

BMI body mass index, *COPD* chronic obstructive pulmonary disease, *DM2* type 2 diabetes mellitus, *DPP4i* dipeptidyl peptidase-4 inhibitors, *GLP1* glucagon-like peptide-1, *HbA1c* glycated hemoglobin, *MDRD* Modification of Diet in Renal Disease, *SD* standard deviation, *SGLT2i* sodium-glucose cotransporter type 2 inhibitors

reached the HbA1c goal 42 days earlier (95% CI 34–49) than non-adherent patients. A similar effect was observed in combination therapies with sulfonylureas 54 (95% CI 44–64) days earlier, GLP1 66 days (9–122), DPP4i 54 days (40–68), and other hypoglycemic agents 50 days (18–82). In contrast,

the time to HbA1c control was not statistically different in patients with the insulin and SGLT2i combination treatment.

We also performed the main analysis using only adherent patients and did not find significant changes. As seen with the entire population, the adjusted difference of RMST was

Table 2 Effects of second-line combination therapies on HbA1c control, adherence, and weight loss

	Overall	Met-formin + sulfonylurea	Met-formin + GLP1	Met-formin + DPP4i	Met-formin + insulin	Met-formin + SGLT2i	Metformin + others
<i>N</i>	28,425	13,216	290	12,000	952	837	1130
HbA1c changes							
Absolute decrease from baseline [mean (SD)]	− 1.82 (1.44)	− 1.84 (1.37)	− 2.13 (1.30)	− 1.66 (1.30)	− 3.90 (2.24)	− 1.64 (1.36)	− 1.73 (1.32)
At 180 days [mean (SD)], missing 28.2%	− 1.66 (1.47)	− 1.65 (1.40)	− 1.99 (1.34)	− 1.51 (1.33)	− 3.76 (2.29)	− 1.54 (1.37)	− 1.55 (1.36)
≥ 0.5 decrease at 180 days (%)	17,312 (60.9)	8229 (62.3)	206 (71.0)	7018 (58.5)	676 (71.0)	505 (60.3)	678 (60.0)
Reached HbA1c goal (%)	18,594 (65.4)	8702 (65.8)	236 (81.4)	7854 (65.4)	587 (61.7)	501 (59.9)	714 (63.2)
Time to HbA1c control (≤ 7%)							
Days to HbA1c normalization (median [IQR])	362.00 [187.00, 772.00]	376.50 [186.00, 868.00]	235.50 [160.00, 378.75]	349.00 [188.00, 700.00]	342.00 [186.00, 770.50]	338.00 [176.00, 595.00]	427.50 [203.00, 907.25]
Median survival time from K–M curves (95% CI)	443 (430–457)	463 (441–486)	245 (228–270)	425 (411–441)	365 (38.3)	476 (413–538)	567 (504–637)
RMST truncated at 720 days (95% CI)	–	455 (451–459)	330 (304–357)	447 (442–452)	452 (435–469)	457 (440–475)	483 (469–498)
Duration of treatment							
Days exposed to initial metformin (median [IQR])	701.0 [375.0, 1156.0]	630.0 [335.0, 1051.0]	560.5 [288.8, 973.0]	812.0 [430.0, 1252.3]	554.5 [307.0, 1042.5]	741.0 [361.0, 1293.0]	698.00 [377.25, 1116.75]
At least 1 year (%)	21,492 (75.6)	9475 (71.7)	196 (67.6)	9668 (80.6)	662 (69.5)	626 (74.8)	865 (76.5)
Days exposed to second-line treatment (median [IQR])	679.0 [362.0, 720.0]	733.0 [380.0, 720.0]	512.0 [300.0, 720.0]	666.0 [364.0, 720.0]	557.0 [279.3, 720.0]	423.0 [219.0, 720.0]	620.5 [315.0, 720.0]
At least 1 year (%)	21,235 (74.7)	10,150 (76.8)	190 (65.5)	8989 (74.9)	630 (66.2)	482 (57.6)	794 (70.3)
Adherence							
MPR [mean (SD)]	87.76 (23.57)	85.75 (24.46)	80.04 (31.09)	92.23 (19.72)	70.17 (30.59)	85.39 (26.78)	82.37 (26.76)
Adherents, MPR ≥ 80% (%)	22,804 (80.2)	10,143 (76.7)	204 (70.3)	10,540 (87.8)	465 (48.8)	646 (77.2)	806 (71.3)

Table 2 (continued)

	Overall	Met-formin + sulfonylurea	Met-formin + GLP1	Met-formin + DPP4i	Met-formin + insulin	Met-formin + SGLT2i	Metformin + others
Impact on weight (kg)							
Absolute difference at 720 days [mean (SD)], missing 52.3%	−4.10 (5.84)	−3.86 (5.61)	−10.40 (11.53)	−4.17 (5.57)	−3.12 (6.80)	−6.83 (7.18)	−3.74 (5.63)
Reduced > 5% in weight (%), missing 52.3%	5370 (39.6)	2434 (37.4)	82 (64.1)	2245 (41.0)	171 (35.3)	233 (56.7)	205 (36.8)
Reduction > 5% of weight compared to metformin + sulfonylurea (<i>N</i> = 13,556)							
aOR (95% CI) ^b	Ref		2.63 (1.8–3.84) ^a	1.28 (1.19–1.38) ^a	0.95 (0.77–1.17)	2.27 (1.85–2.8) ^a	0.94 (0.79–1.14)

BMI body mass index, *CI* confidence interval, *DM2* type 2 diabetes mellitus, *DPP4i* dipeptidyl peptidase-4 inhibitors, *GLP1* glucagon-like peptide-1, *HbA1c* glycated hemoglobin, *IQR* interquartile interval, *K-M* Kaplan–Meier, *MPR* medication possession ratio, *OR* odds ratio, *ref* reference, *RMST* restricted mean survival time, *SD* standard deviation, *SGLT2i* sodium-glucose cotransporter type 2 inhibitors

^aSignificant ($p < 0.001$)

^bAdjusted by days under second-line treatment, baseline weight, days exposed to initial metformin treatment, years with DM2, age, smoking habits, weight, CKDEPI, cancer, ischemic heart disease, chronic respiratory failure, and obesity

significantly different in all therapies except for SGLT2i. The adherence is reducing the time to reach the HbA1c goal but is not changing the results in the comparison with sulfonylureas (Table 3).

3.4 Secondary Outcomes

During the first 2 years, we observed a weight loss (reduction > 5% respect to baseline) in 5370 (39.6%) of the patients. The combination of metformin and a sulfonylurea reached this reduction in 2434 (37.4%) patients, which was the same for other hypoglycemic agents and insulin combinations, although the percentages were higher in the other treatments. The odds of losing more than 5% of weight were significantly higher in combinations with GLP1, OR = 2.63 (95% CI 1.8–3.84), DPP4i, OR = 1.28 (1.19–1.38), and SGLT2i, OR = 2.27 (1.85–2.8) (Table 2).

The duration of previous single metformin treatment was a median 701 (IQR = 375, 1156) days and 75.6% were treated during at least 1 year before the addition. No significant differences between treatments were observed (Table 2). Regarding the exposition period after the second-line therapy, the median number of days was, overall, 679 days (IQR = 362, 720). The sulfonylurea prescriptions were longer, 733 days (IQR = 380, 720), while patients receiving SGLT2i maintained the combined treatment for less time compared with the other treatments, 423 days (IQR = 219, 720).

The average impact on weight was also positive regardless of treatment (Table 2). The overall mean weight reduction at 720 days was 4.10 kg (SD = 5.84) higher in GLP1 and SGLT2i combinations with a mean reduction of 10.40 (SD = 11.5) and 6.83 (SD = 7.2) kg, respectively.

4 Discussion

We analyzed the impact of second-line treatments in HbA1c and weight in 28,425 patients with diabetes and compared them with the combination of metformin and sulfonylureas. In our population, patients receiving metformin and any combination therapy achieved control for HbA1c in less than a year (in median), and GLP1 and insulin combinations reached the control earlier than the combination with sulfonylureas.

The effect of adherence on HbA1c control was also measured for all type of combinations comparing adherent versus non-adherent patients. For all groups except insulins and SGLT2i, adherent patients reached the HbA1c goal earlier than non-adherent patients. Overall, the adherence to antidiabetic drugs was high, which could be explained by the patient perception of disease as we focused on patients with HbA1c uncontrolled DM2 [32]. This was not true with insulins, which includes rapid/short- and intermediate-acting insulins, and their use differs from long-acting insulins. Not having a fixed-dose use is underestimating the measure of adherence [33].

	Median survival (CI 95%)*	RMST (CI 95%)*	Difference RMST (CI 95%)	P-value	Adjusted difference RMST** (CI 95%)	P-value
Metformin+Sulfonylurea	463 (441 to 486)	455 (451 to 459)	ref	-	ref	-
Metformin+GLP1	245 (228 to 270)	330 (304 to 357)	-120.91 (-147.02, -94.81)	<.001	-126.26 (-152.49, -100.03)	<.001
Metformin+DPP4i	425 (411 to 441)	447 (442 to 452)	-0.80 (-7.05, 5.44)	0.801	14.81 (8.61, 21.01)	<.001
Metformin+Insulin	446 (382 to 537)	452 (435 to 469)	-80.3 (-99.54, -61.07)	<.001	-69.18 (-88.07, -50.3)	<.001
Metformin+SGLT2i	476 (413 to 538)	457 (440 to 475)	7.49 (-10.60, 25.58)	0.417	5.42 (-12.32, 23.16)	0.549
Metformin+others	567 (504 to 637)	483 (469 to 498)	31.08 (16.15, 46.01)	<.001	39.01 (24.08, 53.95)	<.001

* From the Kaplan-Meier curves

** Adjusted by Days under 2nd-line treatment, Baseline HbA1c, Days under previous metformin treatment, Years with DM2, Age, Smoking habits, Weight, CKDEPI, Cancer, Ischemic heart disease, Chronic respiratory failure and Obesity

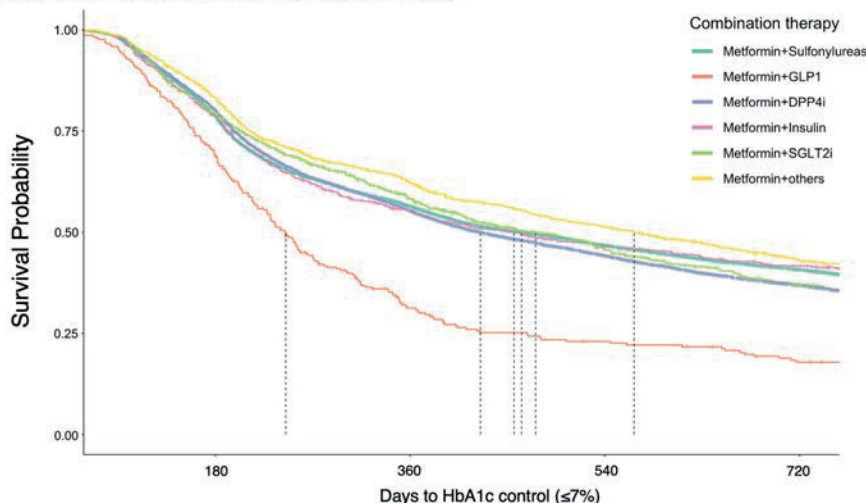


Fig. 2 Kaplan-Meier and restricted mean survival time (RMST) for the time to glycated hemoglobin (HbA1c) control ($\leq 7\%$) by combination treatment over 2 years. CI confidence interval, DM2

type 2 diabetes mellitus, DPP4i dipeptidyl peptidase-4 inhibitors, GLP1 glucagon-like peptide-1, ref reference, SGLT2i sodium-glucose cotransporter type 2 inhibitors

In adherent patients, all combinations showed an earlier achievement of the HbA1c goal except for the SGLT2i group compared to sulfonylureas. However, this combination together with GLP1 and DPP4i groups had higher probabilities to reduce more than 5% of the weight.

4.1 Bibliography Comparison

Vlacho et al. found a trivial effect in HbA1c reductions between groups [30]. Although in our study comparative effect sizes have not been assessed, we highlighted that all groups, except for the SGLT2i, reached HbA1c control faster than those receiving the sulfonylurea combination. The N from our study (43,068) differs drastically from the study by Vlacho et al. (75,808) probably owing to differences in the study design and periods selected, as our study analyses more recent data, and this could explain the discordances

found. The proportion of patients who reached the HbA1c goal was higher in our study but the proportion of patients with a decrease ≥ 0.5 at 180 days was lower. In both studies, the sulfonylurea group had a higher proportion of users reaching the target of HbA1c below 7%. Days to HbA1c normalization were similar in both studies, and significant weight reductions were found for DPP4i and SGLT2i groups [30].

Regarding adherence to second-line therapy, Sicras-Mainar et al. conducted an analysis in 2008–9 but this was not comparable to our study because of the population characteristics, methods, and differences between period studies [31]. In Vlacho et al., they also found high MPR levels but lower than in our study [3]. The sulfonylurea group had the lowest proportion of adherent patients (MPR $> 80\%$), and the time exposed to second-line treatment or persistence in our study was higher in all groups. These differences can be related

Table 3 RSMT to HbA1c control by adherence and second-line combination therapy

Adherent vs non-adherent patients (<i>N</i> = 28,189)	RMST (95% CI) ^a		Adjusted diff RMST (95% CI) ^b	<i>P</i> -value
	Adherents	Non-adherents		
All patients	438 (434–441)	507 (500–514)	–41.89 (–49.34, –34.45)	<0.001
Metformin + sulfonylurea (<i>N</i> = 13,117)	437 (432–442)	511 (502–520)	–54.29 (–64.43, –44.14)	<0.001
Metformin + GLP1 (<i>N</i> = 289)	306 (277–336)	388 (335–442)	–65.6 (–122.1, –9.09)	0.023
Metformin + DPP4i (<i>N</i> = 11,901)	437 (432–442)	520 (508–533)	–54.09 (–67.71, –40.47)	<0.001
Metformin + insulin (<i>N</i> = 938)	456 (432–479)	454 (430–477)	5.15 (–28.68, 38.98)	0.765
Metformin + SGLT2i (<i>N</i> = 830)	445 (425–465)	491 (455–527)	–21.31 (–61.02, 18.41)	0.293
Metformin + others (<i>N</i> = 1114)	468 (451–486)	523 (496–550)	–49.73 (–81.85, –17.61)	0.002
Only adherent patients (<i>N</i> = 22,804)	<i>N</i>	RMST (95% CI) ^a	Adjusted diff RMST (95% CI) ^b	<i>P</i> value
Metformin + sulfonylurea	10,143	438 (433–443)	Ref	–
Metformin + GLP1	204	305 (276–335)	–129.73 (–159.46, –100.01)	<0.001
Metformin + DPP4i	10,540	437 (432–442)	19.5 (12.67, 26.32)	<0.001
Metformin + insulin	465	455 (431–479)	–63.63 (–89.16, –38.1)	<0.001
Metformin + SGLT2i	646	446 (426–466)	12.59 (–7.79, 32.97)	0.226
Metformin + others	806	468 (450–485)	36.97 (19.5, 54.43)	<0.001

CI confidence interval, *diff* difference, *DM2* type 2 diabetes mellitus, *DPP4i* dipeptidyl peptidase-4 inhibitors, *GLP1* glucagon-like peptide-1, *HbA1c* glycated hemoglobin, *RMST* restricted mean survival time, *SGLT2i* sodium-glucose cotransporter type 2 inhibitors

^aEstimation from Kaplan–Meier curves

^bAdjusted by days under second-line treatment, baseline Hb1Ac, days exposed to initial metformin treatment, years with DM2, age, smoking habits, weight, CKDEPI, cancer, ischemic heart disease, chronic respiratory disease, and obesity

to divergences in the methodology but, independently on how this was measured, a higher adherence was associated with greater glycemic control [34–36], lower ratios on some micro-macrovascular outcomes [37], and with lower health-care costs [38, 39].

In our database, adherent patients reduced the time to reach HbA1c control with the GLP1 combination followed by insulins. This is in line with the clinical guidelines, as GLP1 is indicated for an uncontrolled obese population with diabetes while insulins are for general uncontrolled patients with diabetes [15].

4.2 Limitations and Strengths

Among the limitations, the GLP1 group in our context is a pharmacological treatment indicated in combination with metformin and reimbursed only for patients with obesity in the Spanish Health Service. It is a non-comparable group because of the selecting bias, but our aim was to show the general practice in our context. Insulin-treated patients may not be comparable either, as they are usually patients with poor glycemic control. The sulfonylurea group is the second-line gold standard treatment clinically, thus it was chosen as the reference group for the statistical comparison [15]. Another limitation of our study is that we were unable to consider hypoglycemic events because of poor recording in the Catalonia primary care system. Given the

high prevalence of sulfonylureas in our study population, which are known to carry a risk of hypoglycemia, it is important to interpret our results with this limitation in mind. The strengths of our study are the large number of patients, complete sociodemographic data, and adherence to each treatment group.

We have used the smooth algorithm to determine the electronic drug records. It allowed us to automate the data management step and model different aspects of drug exposure. With the algorithm, we obtained all patients who had received the second-line combination treatment, determined treatment initiation, its duration, possible treatment changes or interruptions, and calculated the adherence.

Adherence in databases like SIDIAP is difficult to assess. We have pharmacy invoice records that are not linked to electronic prescriptions, and extra work is required to link both databases. We described a process to easily calculate an approximation to the MPR that is consistent with the official data reported in our region but different to the MRP levels reported by Vlachos et al. [3]. This is a common example of how, in pharmacoepidemiologic studies with similar characteristics, each research team has a unique method to assess drug exposure that may induce biases and less comparable results. We tried to overcome this by using an automatic algorithm that brings additional value to our study [23]. Future research should include a longer follow-up time and focus the analysis on

adverse effects such as bleeding or ictus after the addition of second-line treatment.

5 Conclusions

Second-line treatments reached the HbA1c goal for most of the patients and GLP1 and insulin combinations reached the control earlier than the combination with sulfonylureas. Adherence had a significant effect on reducing the time to HbA1c control in all combinations except for SGLT2i. Weight reduction was also achieved earlier in GLP1 therapy and the SGLT2i combination was able to achieve significant reductions in weight.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40801-023-00374-2>.

Acknowledgements We acknowledge the researchers who participated in this study.

Declarations

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest Dan Ouchi, Carles Vilaplana-Carnerero, Ramon Monfà, Maria Giner-Soriano, Ana Garcia-Sangenís, Ferran Torres, and Rosa Morros have no conflicts of interest that are directly relevant to the content of this article.

Ethics approval The study was approved by the Primary Care Research and Ethics Committee of the Research Institute IDIAP Jordi Gol, Barcelona, Spain (reference number, 21/026-P, 24 February, 2021).

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material Not applicable.

Code availability Not applicable.

Author contributions All authors contributed to the study design. R.Mon., C.V., A.G., M.G.S., and R.Mor. conceived the study. M.G.S., R.Mon., C.V., and D.O. collected the data. D.O. performed the statistical analyses. D.O., A.G., R.Mon., C.V., M.G.S., and R.Mor. were involved in the interpretation of the data. D.O. and C.V. wrote the manuscript and designed the tables and figures. All authors critically revised the report and approved the final version to be submitted for publication. The corresponding author confirms that he had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative

Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

References

- Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al KJ. Epidemiology of type 2 diabetes: global burden of disease and forecasted trends. *J Epidemiol Glob Health*. 2020;10(1):107.
- Turner RC, Cull CA, Frighi V, Holman RR. Glycemic control with diet, sulfonylurea, metformin, or insulin in patients with type 2 diabetes mellitus: progressive requirement for multiple therapies (UKPDS 49). *Endocrinologist*. 1999;9(6):503.
- Vlacho B, Mata-Cases M, Mundet-Tudurí X, Vallès-Callol JA, Real J, Farre M, et al. Analysis of the adherence and safety of second oral glucose-lowering therapy in routine practice from the Mediterranean area: a retrospective cohort study. *Front Endocrinol (Lausanne)*. 2021;12(Jul):1–8.
- Lind M, Pivodic A, Cea-Soriano L, Nerman O, Pehrsson NG, Garcia-Rodriguez LA. Changes in HbA1c and frequency of measuring HbA1c and adjusting glucose-lowering medications in the 10 years following diagnosis of type 2 diabetes: a population-based study in the UK. *Diabetologia*. 2014;57(8):1586–94.
- Huang ES, Davis AM. Glycemic control in older adults with diabetes mellitus. *JAMA*. 2015;314:1509–10.
- Care D, Supp SS. 6. Glycemic targets: standards of medical care in diabetes: 2022. *Diabetes Care*. 2022;45(1):83–96.
- Type 2 diabetes in adults: management. London: National Institute for Health and Care Excellence (NICE); 2022. (NICE Guideline, No. 28.) Available from: <https://www.ncbi.nlm.nih.gov/books/NBK553486/>.
- Mata M, Cos FX, Morros R, Diego L, Barrot J, Berengué M, et al. Pautes per a l'harmonització del tractament farmacològic de la diabetis mellitus tipus 2. Barcelona: Servei Català de la Salut. Departament de Salut. Generalitat de Catalunya; 2017. (Programa d'harmonització farmacoterapèutica de medicaments en l'àmbit de l'atenció primària i comunitària del Servei Català de la Salut; 01/2017).
- Xu Y, Yang Z, Lin H, Shen P, Wang H, Zhan S. Long-term patterns of antidiabetic medication use in patients with type 2 diabetes. *Med Sci Monit*. 2018;24:8707.
- Care D, Suppl SS. 9. Pharmacologic approaches to glycemic treatment: standards of medical care in diabetes: 2022. *Diabetes Care*. 2022;45(Jan):S125–43.
- Mamza J, Mehta R, Donnelly R, Idris I. Important differences in the durability of glycaemic response among second-line treatment options when added to metformin in type 2 diabetes: a retrospective cohort study. *Ann Med*. 2016;48(4):224–34.
- McIntosh B, Cameron C, Singh SR, Yu C, Dolovich L, Houlden R. Choice of therapy in patients with type 2 diabetes inadequately controlled with metformin and a sulphonylurea: a systematic review and mixed-treatment comparison meta-analysis. *Open Med*. 2012;6(2):e62.
- New Drugs for Type 2 Diabetes: Second-Line Therapy — Science Report [Internet]. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health; 2017. (CADTH Therapeutic Review, No. 4.1b.) Available from: <https://www.ncbi.nlm.nih.gov/books/NBK531904/>.

14. Shin, JI. Second-line glucose-lowering therapy in type 2 diabetes mellitus. *Curr Diab Rep.* 2019;19:54. <https://doi.org/10.1007/s11892-019-1171-0>.
15. Fau E, Mata-Cases M, Morros R, Pellicer MA, Ricart W, Vallès JA, et al. Pautes per al tractament farmacològic de la diabetis mellitus tipus 2. Barcelona: Agència de Qualitat i Avaluació Sanitàries de Catalunya; 2013. <https://hdl.handle.net/11351/1403>.
16. Qian D, Zhang T, Zheng P, Liang Z, Wang S, Xie J, et al. Comparison of oral antidiabetic drugs as add-on treatments in patients with type 2 diabetes uncontrolled on metformin: a network meta-analysis. *Diabetes Ther.* 2018;9(5):1945–58.
17. Polonsky WH, Henry RR. Poor medication adherence in type 2 diabetes: recognizing the scope of the problem and its key contributors. *Patient Prefer Adherence.* 2016;10:1299–306.
18. WHO. ICD-10 version. Geneva: WHO; 2019.
19. Anonymous. WHO ATC/DDD Index 2020. WHO Collaborating Centre for Drugs Statistics Methodology. 2020.
20. Bolibar B, Fina Avilés F, Morros R, Garcia-Gil Mdel M, Hermsilla E, Ramos R, Rosell M, Rodríguez J, Medina M, Calero S, Prieto-Alhambra D; Grupo SIDIAP. Base de datos SIDIAP: la historia clínica informatizada de Atención Primaria como fuente de información para la investigación epidemiológica [SIDIAP database: electronic clinical records in primary care as a source of information for epidemiologic research]. *Med Clin (Barc).* 2012;138(14):617–21. Spanish. <https://doi.org/10.1016/j.medcli.2012.01.020>.
21. CatSalut. Servei Català de la Salut. Conjunt mínim bàsic de dades (CMBD). 2022. <http://catsalut.gencat.cat/ca/proveïdors-professionals/registres-cataleg/registres/cmbd/>. Accessed 14 July 2022.
22. Catalan Agency for Health Quality and Evaluation (AQuAS). Updated SARS-CoV-2 data. 2022. <https://aquas.gencat.cat/>. Accessed 14 July 2022.
23. Ouchi D, Giner-Soriano M, Gómez-Lumbreras A, Vedia Urgell C, Torres F, Morros R. SMOOTH algorithm: an automatic method to estimate the most likely drug combination in electronic health records. Development and validation study. *JMIR Preprints.* 2022.
24. Kloeker DE, Davies MJ, Khunti K, Zaccardi F. Uses and limitations of the restricted mean survival time: Illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes. *Ann Intern Med.* 2020;172(8):541–52.
25. A'Hern RP. Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials? *J Clin Oncol.* 2016;34(28):3474–6.
26. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med.* 2011;30(19):2409–21.
27. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics.* 2014;15(2):222–33.
28. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Development Core Team; 2017.
29. Dessau RB, Pipper CB. “R”: project for statistical computing. Ugeskr Laeger. 2008.
30. Vlachos B, Mundet-Tuduri X, Mata-Cases M, Vallès-Callol JA, Real J, Farré M, et al. Analysis of the effectiveness of second oral glucose-lowering therapy in routine clinical practice from the Mediterranean area: a retrospective cohort study. *Diabetes Res Clin Pract.* 2021;171:1–10.
31. Sicras-Mainar A, Font-Ramos B, Roldán-Suárez C, Navarro-Artieda R, Ibáñez-Nolla J. Caracterización y costes asociados al perfil del paciente con diabetes tipo 2 en tratamiento con metformina al que se le añade un segundo fármaco antidiabético oral: estudio de base poblacional. *Endocrinol Nutr.* 2013;60(10):557–69.
32. Gil-Girbau M, Aznar-Lou I, Peñarubia-María MT, Moreno-Peral P, Fernández A, Bellón JA, et al. Reasons for medication non-initiation: a qualitative exploration of the patients' perspective. *Res Soc Admin Pharm.* 2020;16(5):663–72.
33. Sarbacker GB, Urteaga EM. Adherence to insulin therapy. *Diabetes Spectr.* 2016;29(3):166–70.
34. Kim N, Agostini JV, Justice AC. Refill adherence to oral hypoglycemic agents and glycemic control in veterans. *Ann Pharmacother.* 2010;44(5):800–8. <https://doi.org/10.1345/aph.1M570>.
35. Ibrahim AO, Agboola SM, Elegbede OT, Ismail WO, Agbesanwa TA, Omolayo TA. Glycemic control and its association with sociodemographics, comorbid conditions, and medication adherence among patients with type 2 diabetes in southwestern Nigeria. *J Int Med Res.* 2021. <https://doi.org/10.1177/03000605211044040>.
36. Sánchez-Hernández MS, et al. Impact of adherence to Mediterranean diet and/or drug treatment on glycaemic control in type 2 diabetes mellitus patients: DM2-CUMCYL study. *Prim Care Diab.* 2020. <https://doi.org/10.1016/j.pcd.2020.06.008>.
37. Evans M, Engberg S, Faurby M, Fernandes JDDR, Hudson P, Polonsky W. Adherence to and persistence with antidiabetic medications and associations with clinical and economic outcomes in people with type 2 diabetes mellitus: a systematic literature review. *Diabetes Obes Metab.* 2022;24(3):377–90. <https://doi.org/10.1111/dom.14603>.
38. Balkrishnan R, Rajagopalan R, Camacho FT, Huston SA, Murray FT, Anderson RT. Predictors of medication adherence and associated health care costs in an older population with type 2 diabetes mellitus: a longitudinal cohort study. *Clin Ther.* 2003;25(11):2958–71. [https://doi.org/10.1016/S0149-2918\(03\)80347-8](https://doi.org/10.1016/S0149-2918(03)80347-8).
39. Sokol MC, McGuigan KA, Verbrugge RR, Epstein RS. Impact of medication adherence on hospitalization risk and healthcare cost. *Med Care.* 2005;43(6):521–30. <https://doi.org/10.1097/01.mlr.0000163641.86870.af>.

5. Resumen y Análisis de Resultados del Tratamiento de la T2DM

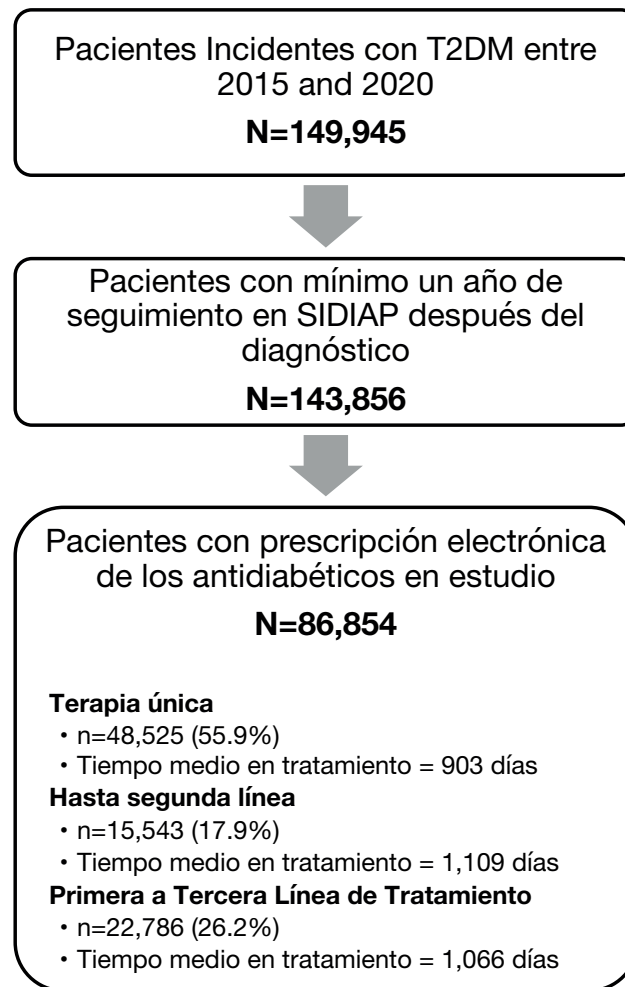
En este capítulo, se exponen y debaten los resultados más destacados de esta tesis doctoral. Principalmente, hemos evidenciado que la implementación de nuevos métodos destinados a mejorar la calidad de los datos de medicamentos en registros electrónicos de salud representa una valiosa contribución al campo de la investigación. A través del algoritmo “smooth”, no solo se automatiza la depuración de datos, sino que también se simplifica el análisis de medicamentos en estudios farmacológicos, lo que permite una modelización más avanzada de los tratamientos a lo largo del tiempo. Para ilustrar el potencial del algoritmo, se llevaron a cabo dos estudios centrados en pacientes con T2DM, cuyos resultados principales se presentan a continuación.

5.1 Descripción Longitudinal de los Tratamientos Antidiabéticos

Tratamiento Inicial

En total, incluimos a 86,854 pacientes en nuestro estudio (Figura 13), de los cuales el 78.3% inició su tratamiento con monoterapia. La metformina fue el fármaco más comúnmente prescrito, 67.4% de los pacientes, una tendencia que concuerda con investigaciones previas [45], [46], [47] y está en línea con las actuales guías para el manejo de la T2DM [48]–[50].

Figura 13. Diagrama de Flujo de los Pacientes en Estudio



En segundo lugar, destacamos que un 4.1% de los pacientes iniciaron su tratamiento con insulina como terapia de primera línea. Este grupo se caracterizó por ser el que comenzó el tratamiento antes y por presentar niveles más elevados de HbA1c en comparación con aquellos que iniciaron con otras monoterapias. Si bien investigaciones previas han mostrado porcentajes aún más altos de pacientes que inician con insulina como tratamiento inicial para la T2DM [45], [51], nuestros valores son similares al porcentaje reportado por Mata-Cases et al. en el período de 2007 a 2018 en SIDIAP [52].

Monoterapia y Terapia Combinada

A pesar de que la monoterapia es la opción recomendada como tratamiento de primera línea [40], en nuestro estudio, el 21.7% (Tabla 7) de los pacientes iniciaron con una terapia combinada (CT). Posiblemente se deba a que estos pacientes presentaban un peor control metabólico en comparación con sus contrapartes que eligieron la monoterapia (más del 73% de ellos tenían $HbA1c \geq 8\%$). De hecho, está en línea con otros estudios realizados en Cataluña [52].

Existen datos que respaldan el uso de la CT desde el inicio para lograr rápidamente los objetivos glucémicos [53] y debería considerarse especialmente en aquellos pacientes cuyos niveles de $HbA1c$ se sitúen de 1.5% a 2% por encima del objetivo [54]. Aunque otros estudios también han descrito la CT como tratamiento de primera línea, no es tan común como en nuestro estudio; por ejemplo, Carney et al. informaron un 6.9% de pacientes tratados inicialmente con CT, mientras que Nicolucci et al. describieron la combinación de metformina y sulfonilureas en el 14.6% de los pacientes recién tratados [47], [55].

Tabla 7. Descripción de los Tratamientos de Primera a Tercera Línea en Función de Edad, $HbA1c$ y Cumplimiento

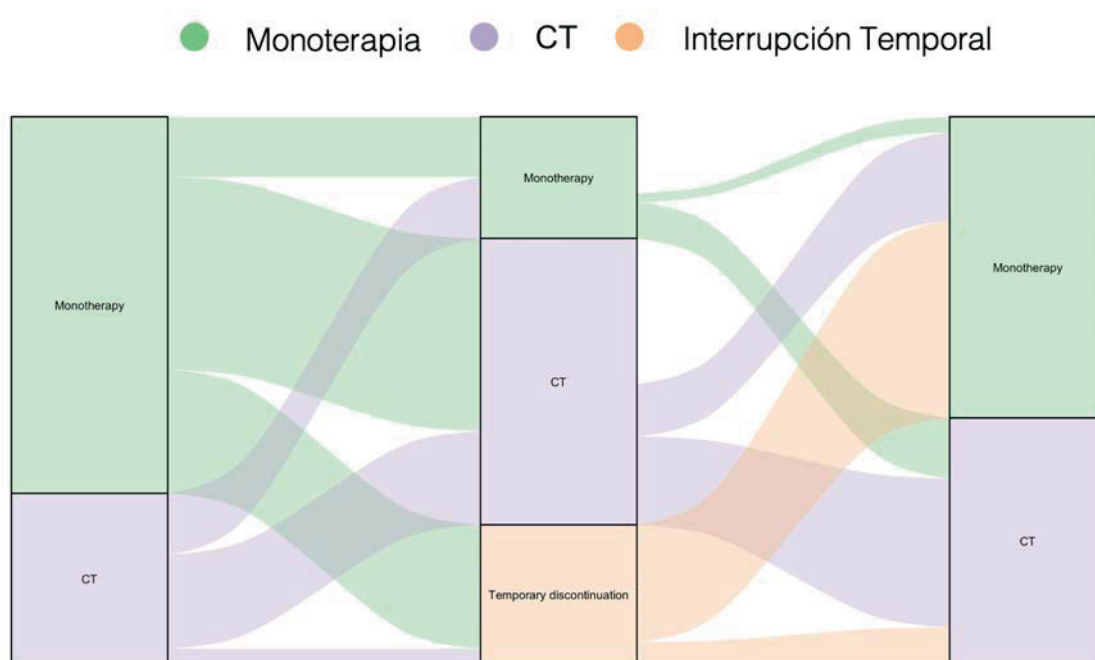
	TERAPIA INICIAL			SEGUNDA LÍNEA				TERCERA LÍNEA			
	Total	>70 años	$HbA1c \geq 8\%$	Total	>70 años	MPR $\geq 80\%*$	$HbA1c \geq 8\%$	Total	>70 años	MPR $\geq 80\%$	$HbA1c \geq 8\%$
n (%)	86854	17589	43177	38329	6788	16624	21883	22786	4049	9124	12755
Mono- terapia	68011 (78.3)	14060 (79.9)	31394 (72.7)	8521 (22.2)	2116 (31.2)	2916 (17.5)	4493 (20.5)	12528 (55.0)	2510 (62.0)	5147 (56.4)	6014 (47.2)
CT	18843 (21.7)	3529 (20.1)	11783 (27.3)	20044 (52.3)	2914 (42.9)	10328 (62.1)	13109 (59.9)	10258 (45.0)	1539 (38.0)	3977 (43.6)	6741 (52.8)
Discontinua- ción**	0 (0.0)	0 (0.0)	0 (0.0)	9764 (25.5)	1758 (25.9)	3380 (20.3)	4281 (19.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

* Adherente al tratamiento anterior. **No hubo seguimiento de la tercera línea de tratamiento.

En lo que respecta a los tratamientos de segunda línea en adelante, dado que la T2DM es principalmente una enfermedad progresiva, las recomendaciones incluyen intensificar el tratamiento, generalmente mediante la adición gradual de medicamentos a la metformina para mantener el objetivo de $HbA1c$ [40], [54], [56]. En nuestro estudio, el 44.1% de los pacientes cambió su primer tratamiento antidiabético por una CT. Sin embargo, observamos un patrón inusualmente frecuente (25.5%) que implicaba una interrupción temporal del tratamiento antidiabético después de recibir una terapia de primera línea, seguida en la mayoría de los casos por un regreso a la monoterapia. También identificamos a 5,401 pacientes (6.2%) que comenzaron con CT y posteriormente redujeron su tratamiento a monoterapia (Figura 14). La reducción del

tratamiento en estos casos no sigue un patrón establecido para la T2DM y probablemente se deba a múltiples factores.

Figura 14. Flujo de Pacientes de Primera a Tercera Línea



Patrones de Tratamiento Longitudinal

Llevamos a cabo un análisis exhaustivo de los patrones longitudinales más frecuentes, teniendo en cuenta la dimensión temporal y estratificando según si iniciaron con monoterapia o con CT (Figura 15).

Entre los pacientes que mantuvieron la misma línea terapéutica, la metformina fue el medicamento más comúnmente utilizado con una duración mediana de 27 meses (IQR 14 a 45) y se inició en un plazo de 6 meses desde el diagnóstico. Además, con respecto a la metformina, observamos un patrón que interrumpía el tratamiento con Metformina a los 6 meses (IQR 3 a 14), durante aproximadamente 4 meses (IQR de 2 a 12) y luego retoma la monoterapia con metformina durante unos 13 meses (IQR de 5 a 25).

Los pacientes que iniciaron con insulina tuvieron una duración mediana de 23 meses (rango intercuartílico de 12 a 42). Tanto los inhibidores de la DPP4 (DPP4i) como la combinación de

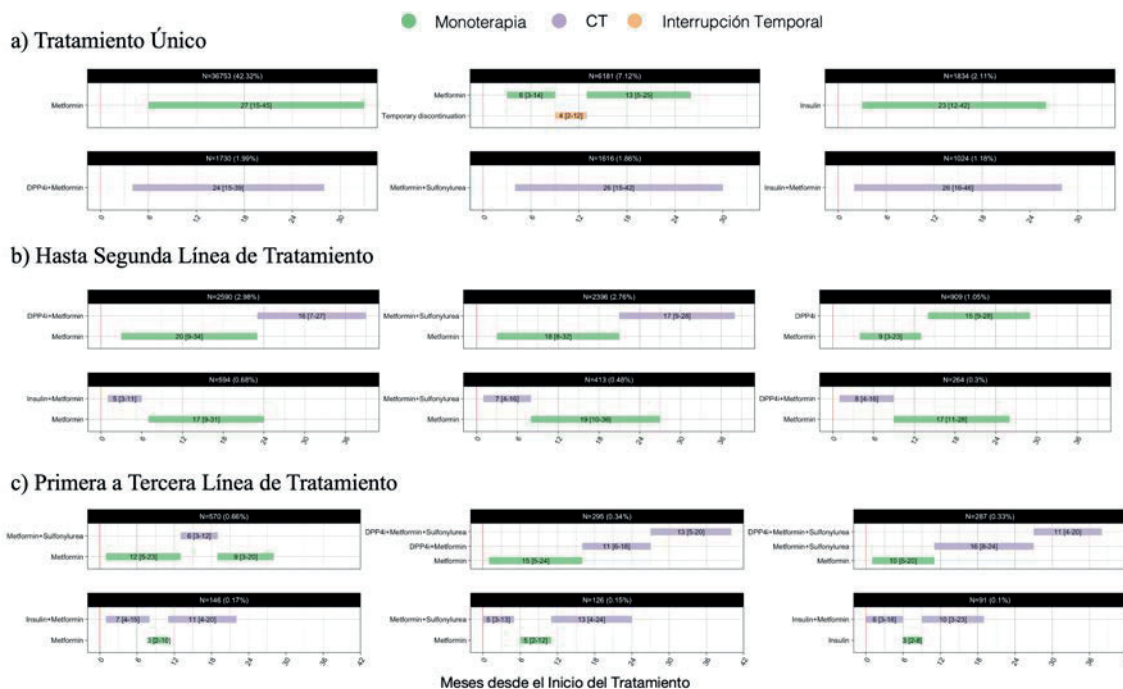
DPP4i con metformina mostraron duraciones similares de tratamiento, 22 meses (IQR de 12 a 35) y 24 meses (IQR 15 a 39), respectivamente (Figura 15a).

La Figura 15b representa los patrones de tratamiento de aquellos pacientes que recibieron dos líneas terapéuticas. El patrón más común implicó comenzar con metformina y luego cambiar a una CT. La CT más frecuente consistió en agregar un DPP4i a la metformina, observado en el 3% de los casos, con una duración mediana de 20 meses (IQR 9 a 34). Aquellos que cambiaron a DPP4i o sulfonilurea en monoterapia tuvieron un período inicial de metformina más corto, 9 meses (IQR 3 a 23) y 6 meses (IQR 6 a 30), respectivamente.

La pauta más frecuente de primera a tercera línea fue comenzar con metformina en monoterapia, agregar un segundo agente antidiabético en segunda línea y luego volver a metformina en monoterapia. De estos, una combinación frecuente fue el uso de DPP4i después de una mediana de 12 meses (IQR 5 a 23) en monoterapia y mantenerse en esta terapia durante 6 meses (IQR 3 a 12), antes de volver a la metformina en monoterapia (Figura 15c). Por otro lado, hubo patrones de terapias en los que los pacientes comenzaron con una combinación de metformina y otro agente antidiabético (DPP4i, sulfonilurea o insulina) durante aproximadamente un año antes de retirarla en la terapia de segunda línea, dejando solo metformina durante 3-5 meses antes de volver al tratamiento inicial (Figura 15c).

Finalmente, identificamos patrones de tratamiento que consistieron en dos adiciones consecutivas. Por ejemplo, el cambio de metformina a una combinación de metformina con un inhibidor de la DPP4 (DPP4i) y luego a una terapia triple de metformina con DPP4i y sulfonilurea (n = 295, 0,34 %).

Figura 15. Patrones de Tratamiento Longitudinal



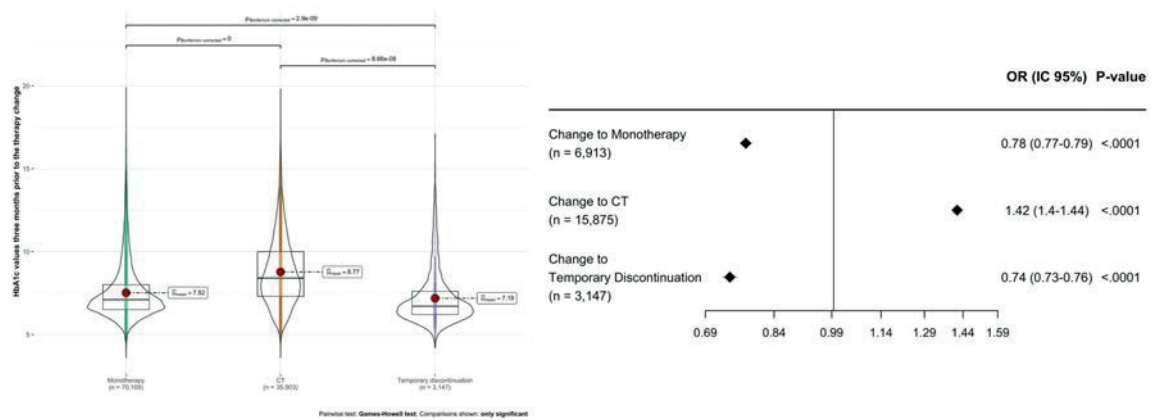
Representación de los tres patrones más frecuentes según el número de cambios y si inicia con monoterapia o con CT. La duración de cada fase se expresa dentro del segmento y representa la mediana y su IQR.

Asociación con la HbA1c

Coincidiendo con otros estudios, como el de Wang [65] y el de Montvida et al. [71], hemos utilizado los niveles de HbA1c durante el período de estudio para analizar su relación con los patrones de tratamiento longitudinal. Nuestros hallazgos claramente indican que cualquier cambio en el patrón de tratamiento está precedido por un cambio en los niveles de HbA1c. Cuando la HbA1c aumenta, observamos adiciones al tratamiento, siguiendo las recomendaciones para mantener los objetivos glucémicos [40]. Por otro lado, cuando la HbA1c disminuye, esto se traduce en un cambio a monoterapia o una interrupción temporal del tratamiento (Figura 16).

Pocos estudios en EHR han examinado longitudinalmente los patrones de tratamiento de la T2DM, incluyendo los cambios de tratamiento a lo largo del tiempo y su asociación con los niveles de HbA1c. Como se observa en la Figura 16, los cambios en el tratamiento suelen ocurrir después de 6 meses o múltiplos de este período (12, 18 y 24 meses), lo que probablemente sea debido a la frecuencia recomendada con la que se realizan las pruebas de HbA1c en Cataluña.

Figura 16. Asociación de los Patrones Longitudinales a Cambios en HbA1c



OR de la asociación entre el cambio de terapia y tener un nivel de HbA1c $\geq 8\%$ en los tres meses anteriores al cambio. Los modelos se ajustaron a parámetros basales.

Otras posibles causas de cambio de tratamiento pueden ser los efectos secundarios o la falta de adherencia al tratamiento que podría resultar en un aumento de la HbA1c y, en consecuencia, en la modificación del tratamiento. En nuestro estudio, se encontró que el 41% de los pacientes tratados con metformina tenían una adherencia subóptima (MPR $< 80\%$), lo que sugiere que, en última instancia, es la HbA1c la que guía los ajustes terapéuticos.

5.2 Impacto del Tratamiento de Segunda Línea en la HbA1c

Descripción de la Población

En total, se analizó una población de 28,425 pacientes. La edad promedio fue de 63.63 años (SD = 11.38), de los cuales el 38.1% eran mujeres. La media de peso en la población se situó en 83.64 kg (SD = 16.64). Entre los pacientes que recibieron terapia combinada con GLP1, se observó una mayor proporción de mujeres, el 52.1%. Además, este grupo mostró un peso promedio más elevado de 107.26 kg (SD = 20.52) y fue diagnosticado aproximadamente 3.33 años antes de la adición (SD = 3.99). Las características clínicas fueron similares en los diferentes grupos, con una tasa media de filtración glomerular, calculada mediante CKDEPI, de 81.12 (SD = 12.92) mL/min/1.73 m². Las principales comorbilidades en la población incluyeron hipertensión, que afectaba al 69.9% de los pacientes, obesidad, presente en el 64.4%, y dislipidemia, detectada en el 62.7% de los casos (más detalle en Tabla 8).

Tabla 8. Características Demográficas y Clínicas de la Población

	GLOBAL	METF+ SULFONYLUREA	METFORMIN+ GLP1	METF+ DPP4I	METF+ INSULIN	METF+ SGLT2I
N	28425	13216	290	12000	952	837
HbA1c a Fecha Índice (media (SD))	8.63 (1.34)	8.64 (1.28)	8.57 (1.31)	8.46 (1.18)	10.83 (2.09)	8.54 (1.28)
Años en T2DM (media (SD))	5.88 (4.64)	5.77 (4.53)	3.33 (3.99)	6.14 (4.70)	5.01 (4.72)	5.10 (4.35)
Mujer (%)	10831 (38.1)	4956 (37.5)	151 (52.1)	4588 (38.2)	359 (37.7)	326 (38.9)
Edad (media (SD))	63.63 (11.38)	62.64 (11.10)	54.92 (10.87)	64.98 (11.22)	62.81 (13.47)	60.10 (10.68)
Hábito Fumador (%)						
<i>No Fumador</i>	12195 (42.9)	5579 (42.2)	99 (34.1)	5310 (44.2)	349 (36.7)	330 (39.4)
<i>Ex Fumador</i>	4997 (17.6)	2384 (18.0)	61 (21.0)	1992 (16.6)	203 (21.3)	178 (21.3)
<i>Fumador</i>	11233 (39.5)	5253 (39.7)	130 (44.8)	4698 (39.1)	400 (42.0)	329 (39.3)
Peso (media (SD))	83.64 (16.64)	84.13 (16.38)	107.26 (20.52)	82.34 (16.01)	81.09 (18.23)	90.53 (18.43)
BMI (media (SD)), <i>Missings 52.4%</i>	31.39 (5.30)	31.44 (5.17)	39.81 (6.45)	31.02 (5.07)	30.42 (5.98)	34.54 (5.96)
<i>BMI≥30 (%)</i>	7626 (26.8)	3705 (28.0)	123 (42.4)	2949 (24.6)	240 (25.2)	315 (37.6)
CKDEPI (media (SD))	81.12 (12.92)	82.12 (12.02)	83.87 (12.22)	80.38 (13.25)	79.39 (15.07)	83.24 (10.92)
MDRD (media (SD)), <i>Missings 0.8%</i>	59.19 (3.52)	59.40 (2.98)	59.06 (4.18)	59.11 (3.62)	58.51 (5.03)	59.58 (2.57)
Comorbilidades, N (%)						
Ictus isquémico	443 (1.6)	212 (1.6)	1 (0.3)	176 (1.5)	17 (1.8)	9 (1.1)
Aterosclerosis	1994 (7.0)	909 (6.9)	12 (4.1)	836 (7.0)	97 (10.2)	51 (6.1)
Enfermedad cerebrovascular	1644 (5.8)	750 (5.7)	17 (5.9)	698 (5.8)	65 (6.8)	36 (4.3)
Cáncer	4351 (15.3)	1946 (14.7)	33 (11.4)	1907 (15.9)	165 (17.3)	104 (12.4)
Cardiopatía isquémica	3394 (11.9)	1547 (11.7)	25 (8.6)	1388 (11.6)	137 (14.4)	143 (17.1)
Dislipidemia	17814 (62.7)	8237 (62.3)	171 (59.0)	7617 (63.5)	546 (57.4)	520 (62.1)
Fibrilación auricular	2191 (7.7)	949 (7.2)	23 (7.9)	947 (7.9)	97 (10.2)	51 (6.1)
Hipertensión	19866 (69.9)	9219 (69.8)	208 (71.7)	8365 (69.7)	642 (67.4)	572 (68.3)
Insuficiencia cardíaca	1465 (5.2)	615 (4.7)	17 (5.9)	623 (5.2)	89 (9.3)	44 (5.3)
Insuficiencia respiratoria crónica	3076 (10.8)	1234 (9.3)	21 (7.2)	1419 (11.8)	117 (12.3)	46 (5.5)
EPOC	2464 (8.7)	1114 (8.4)	22 (7.6)	1020 (8.5)	123 (12.9)	77 (9.2)
Obesidad	18311 (64.4)	8753 (66.2)	283 (97.6)	7339 (61.2)	557 (58.5)	658 (78.6)

SD= Desviación Estándar

Resultados Principales

Respecto al análisis principal, se comparó el tiempo hasta el control de la enfermedad con aquellos pacientes que recibieron la combinación de metformina y sulfonilureas como referencia (véase la Figura 17). Se observó que el 50% de los pacientes que fueron tratados con metformina

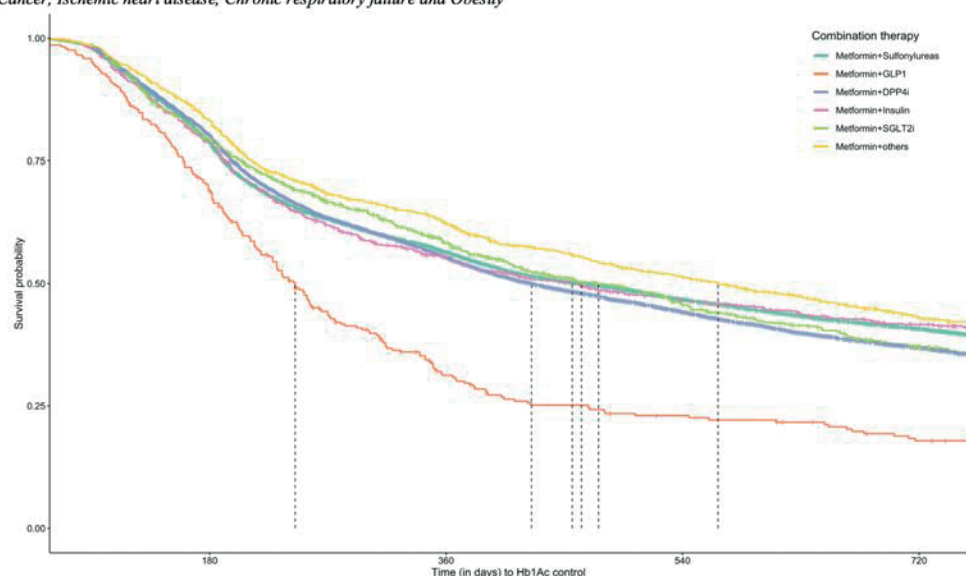
junto con cualquier terapia combinada lograron el control de los niveles de HbA1c en un período de tiempo mediano inferior al año y medio. De los diferentes tratamientos de segunda línea, las combinaciones con GLP1 e Insulina alcanzaron el control de HbA1c antes que la combinación de Metformina con Sulfonilurea. En el caso del GLP1, la diferencia ajustada fue de 126 días, tres meses antes en promedio ($p < .001$).

Figura 17. Kaplan-Meier y RMST para el Tiempo a HbA1c <8%

	Median survival (CI 95%)*	RMST (CI 95%)*	Difference RMST (CI 95%)	P-value	Adjusted difference RMST** (CI 95%)	P-value
Metformin+Sulfonylurea	463 (441 to 486)	455 (451 to 459)	ref	-	ref	-
Metformin+GLP1	245 (228 to 270)	330 (304 to 357)	-120.91 (-147.02, -94.81)	<.001	-126.26 (-152.49, -100.03)	<.001
Metformin+DPP4i	425 (411 to 441)	447 (442 to 452)	-0.80 (-7.05, 5.44)	0.801	14.81 (8.61, 21.01)	<.001
Metformin+Insulin	446 (382 to 537)	452 (435 to 469)	-80.3 (-99.54, -61.07)	<.001	-69.18 (-88.07, -50.3)	<.001
Metformin+SGLT2i	476 (413 to 538)	457 (440 to 475)	7.49 (-10.60, 25.58)	0.417	5.42 (-12.32, 23.16)	0.549
Metformin+others	567 (504 to 637)	483 (469 to 498)	31.08 (16.15, 46.01)	<.001	39.01 (24.08, 53.95)	<.001

* From the Kaplan-Meier curves

** Adjusted by Days under 2nd-line treatment, Baseline HbA1c, Days under previous metformin treatment, Years with DM2, Age, Smoking habits, Weight, CKDEPI, Cancer, Ischemic heart disease, Chronic respiratory failure and Obesity



La línea vertical representa la mediana de supervivencia.

También se exploró el efecto de la adherencia al tratamiento. Analizamos el tiempo hasta el control de HbA1c en todas las combinaciones de segunda línea, comparando los pacientes adherentes (MPR>80%) con aquellos que no lo eran. En todos los grupos, excepto los que recibieron insulinas y SGLT2i, se encontró que los pacientes adherentes alcanzaron el objetivo de HbA1c antes que los no adherentes.

En general, la adherencia a los fármacos antidiabéticos en nuestra muestra fue alta. Este fenómeno podría explicarse por la percepción de la enfermedad por parte de los pacientes, ya que nuestra investigación se enfocó en pacientes con T2DM que no tenían un control adecuado de sus niveles de HbA1c [57]. Sin embargo, es importante señalar que en el caso de las insulinas, que incluyen tanto insulinas de acción rápida/corta como intermedia, la adherencia no fue tan alta. Esto se puede deber a que el manejo de las insulinas implica un enfoque diferente y más personalizado, y no puede medirse de la misma manera que otras terapias [58].

6. Discusión

6.1 El Algoritmo *smooth*

Desde que se popularizó el uso de EHR en estudios farmacoepidemiológicos, los investigadores han estado trabajando en el formas de modelar la exposición a medicamentos a lo largo del tiempo [59]. Tarea que aumenta en complejidad cuando se trata de analizar la exposición a múltiples tratamientos farmacológicos, algo común en personas mayores, quienes a menudo pueden tomar hasta 5 medicamentos para una determinada patología [60].

Comúnmente, abordar este problema implica la aplicación de una serie de criterios o algoritmos definidos por los investigadores, basados en su conocimiento previo de los datos. Aunque estos criterios suelen ser sólidos, tienden a ser altamente específicos y, por lo tanto, poco aplicables en otros contextos de estudio. De hecho, en raras ocasiones encontramos en la literatura métodos que aborden de manera automática y generalizable esta tarea, lo que nos motivó para desarrollar el método.

Algoritmo para Determinar el Tratamiento Más Probable

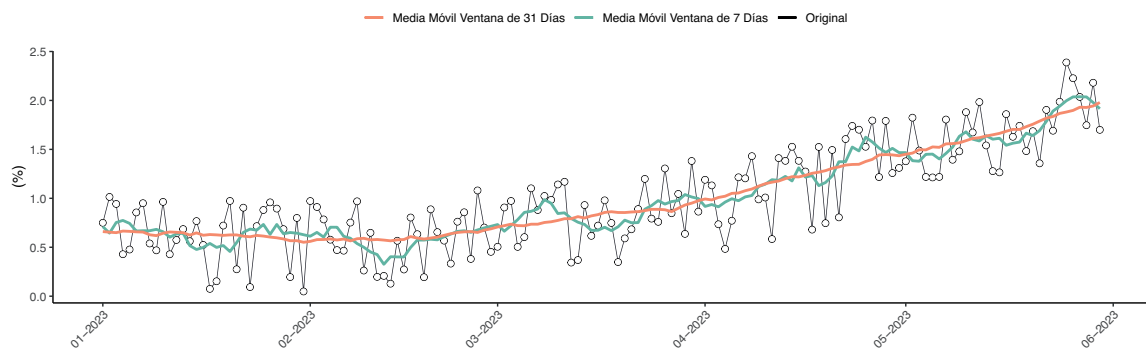
La principal función del algoritmo “smooth” es la de eliminar perfiles de tratamientos irregulares presentes en los datos, que suelen ser complicados de explicar y, sobre todo, de analizar. Para lograrlo, empleamos técnicas estadísticas comunes en el análisis de series temporales para calcular el tratamiento más probable para cada paciente a lo largo del tiempo. En concreto, se usaron el concepto de las medias móviles (MA, por sus siglas en inglés, *Moving Averages*), adaptado al tipo de datos que conforman los registros de medicamentos en EHR.

Las MA son una técnica ampliamente utilizada en series temporales para suavizar los datos y resaltar tendencias o patrones subyacentes. Consiste en calcular el promedio de un conjunto de valores consecutivos en la serie temporal, y sustituirlo por los valores individuales. Aunque existen diversas formas de aplicar las MA, el algoritmo “smooth” se basa en la media móvil simple (SMA, por sus siglas en inglés, *Simple Moving Average*).

La SMA se calcula como el promedio de un determinado número de puntos de datos consecutivos incluidos en una ventana de tiempo. Por ejemplo, si se emplea una ventana de 7 días,

se suman los valores de los últimos 7 días y se divide esa suma entre 7 para obtener el valor promedio de ese punto temporal. Posteriormente, se desplaza la ventana una unidad (normalmente días), y se vuelve a calcular la media móvil sobre los valores incluidos en la ventana. A mayor tamaño de la ventana, mayor será el efecto de suavizado y más se destacan las tendencias por encima del dato original (véase Figura 18).

Figura 18. Medias Móviles en Series Temporales



Serie temporal de valores simulados (en %). Los puntos y línea negra representa el valor original. La serie se muestra suavizada con MA de 7 días (verde) y 31 días (naranja).

En EHR, una de sus características principales es la longitudinalidad de los datos. De esta manera, es posible considerar estas bases de datos como series temporales de variables clínicas, lo que posibilita aplicar estos métodos ampliamente establecidos en el análisis de series temporales para identificar tendencias y suavizar las series.

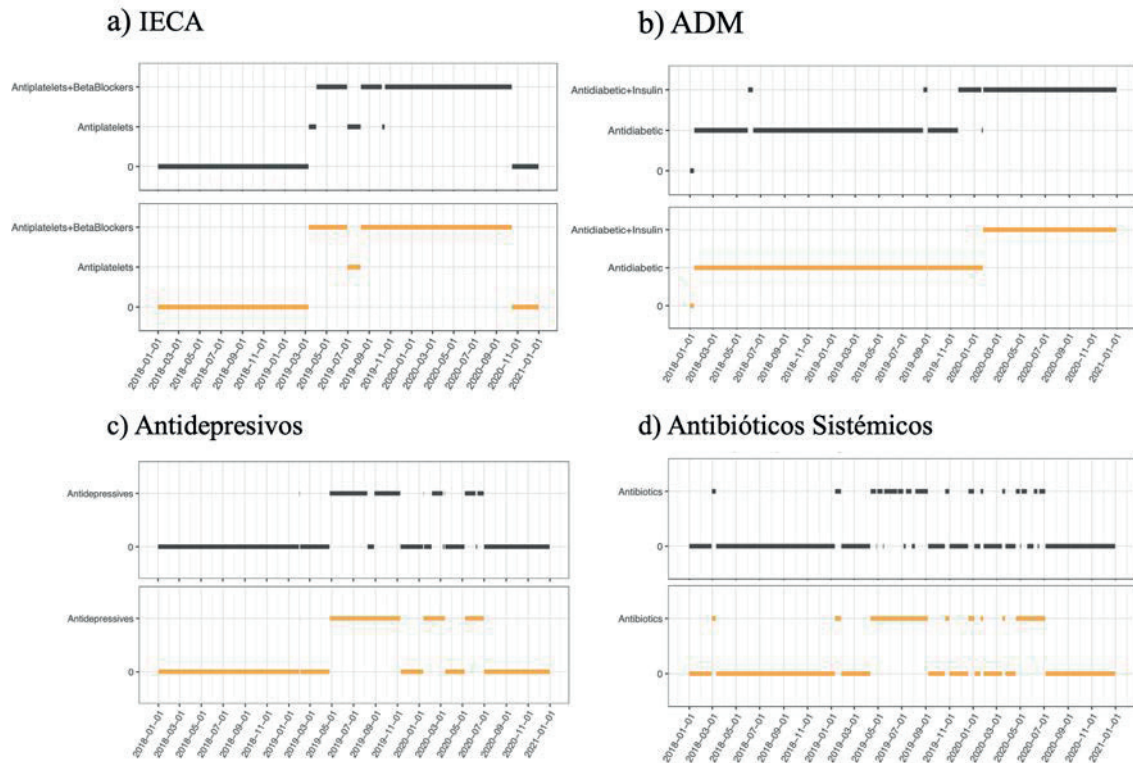
En el caso de los datos de medicamentos, la principal diferencia reside en que estos datos no son numéricos y no se puede hacer el cálculo directo de una media aritmética. En su lugar, sí es posible determinar el tratamiento más frecuente, o moda¹¹.

La Figura 19 ilustra cómo el algoritmo “smooth” actúa en los tratamientos a lo largo del tiempo de cuatro pacientes expuestos a diferentes medicamentos (ejemplos extraídos de la validación, más información en Tabla Sup. 4 del Apéndice 2). A simple vista, se puede observar cómo el algoritmo elimina períodos de tratamiento muy cortos, destacando la tendencia central o

¹¹ En estadística, la moda es el valor que aparece con mayor frecuencia en un conjunto de datos. Enlace: <https://bookdown.org/aquintela/EBE/la-moda-1.html>

más probable. Por ejemplo, en el caso de los pacientes con fibrilación auricular no valvular tratados con IECA, los registros originales (antes de la aplicación del algoritmo) muestran que el paciente comienza con antiagregantes plaquetarios, cambia rápidamente a una combinación de betabloqueantes y luego alterna entre diferentes tratamientos hasta que, a finales de 2019, se mantiene en terapia combinada hasta el final del seguimiento. Sin embargo, el resultado del algoritmo "smooth" simplifica notablemente el perfil longitudinal del paciente. Este comienza en marzo de 2019 directamente con terapia combinada, se le eliminan los betabloqueantes en julio de 2019 y, durante 90 días, recibe tratamiento simple con antiagregantes plaquetarios, para luego volver a la terapia combinada hasta el final del estudio.

Figura 19. Ejemplos de Usar el “smooth” en Diferentes Medicamentos



Ejemplos de cuatro pacientes revisados durante la validación. En negro, el perfil de tratamiento original, y en amarillo el tratamiento suavizado mediante algoritmo.

Para utilizar el algoritmo, es necesario especificar cuál es el grado de suavizado que se desea, o lo que es lo mismo, definir un valor para el parámetro Wt , la ventana de suavizado (en su notación matemática, Wt). Este tipo de parámetro es común en métodos estadísticos similares, como los modelos autorregresivos o medias móviles [61], y el valor que se asigne puede tener un

impacto significativo en los resultados [62], [63], [64]. Teóricamente, si el valor de Wt es pequeño, apenas se esperan cambios en los perfiles de tratamiento longitudinal, pero a medida que el valor del parámetro aumenta, más efecto se espera del algoritmo y más suavizados se mostrarán los tratamientos.

Los resultados de la simulación revelaron que, en el peor de los casos, el algoritmo realiza cambios significativos (>5% de los registros modificados) en el 40% de los pacientes. En rangos de Wt de entre 30 a 90 días, este porcentaje cae entre un 0.7% y un 10.3% (en 60 días es de 2.4%). Cuando Wt es inferior a 30 días, no esperamos cambios significativos (Tabla 9).

Tabla 9. Resultados de la Simulación con Diferentes Valores de Wt

VENTANA Wt , EN DÍAS						
N = 7,132 pacientes	10	20	30	45	60	90
% de cambio* mediana [IQR]	0.27 [0.09, 0.36]	0.46 [0.27, 0.73]	0.64 [0.36, 1.09]	1.00. [0.46, 1.64]	1.46 [0.73, 2.37]	2.28. [1.09, 3.65]
>5%, n (%)	0 (0.0)	0 (0.0)	0 (0.0)	16 (0.7)	58 (2.4)	291 (10.3)

* $100 \times \text{Total de días modificados} / \text{Total de días en exposición}$

Validación del Algoritmo

Para comprender cómo se desempeña el algoritmo desde una perspectiva clínica o farmacológica, se realizó un estudio de validación en cuatro escenarios en los que podría ser necesario el uso del algoritmo: combinación de tres o más medicamentos en la prevención cardiovascular (antiplaquetarios, beta bloqueantes e inhibidores de la ECA (IECA) con ARA-II), combinación a largo plazo de dos grupos farmacológicos para el tratamiento crónico de la T2DM (insulinas y antidiabéticos orales), tratamientos agudos a largo plazo con propensión a la interrupción (antidepresivos) y tratamiento de corta duración (antibióticos sistémicos).

Esta validación se llevó a cabo con la participación de cuatro farmacólogas clínicas independientes, quienes evaluaron en un proceso ciego a un total de 400 pacientes, con 100 pacientes cada una, distribuidos en cuatro grupos de medicamentos diferentes (consultar Tabla Sup 4). La evaluación del algoritmo se realizó cualitativamente, centrándonos en determinar si el algoritmo mejoraba, no alteraba o empeoraba el perfil de tratamiento longitudinal de los pacientes.

Ventajas del Algoritmo

Consideramos que el algoritmo “smooth” tiene un enorme potencial para evaluar la exposición a tratamientos combinados, especialmente en casos de tratamientos crónicos. Esto se debe a que nos permite obtener una secuencia temporal detallada de la exposición a los medicamentos. Esta secuencia nos brinda la capacidad de modelar de manera más precisa la exposición a los fármacos, identificar interrupciones, cambios y periodos de interacción con otros medicamentos de corta duración. Además, con el algoritmo “smooth” podemos calcular el tiempo de exposición de manera sencilla, utilizando únicamente los envases de las recetas electrónicas sin necesidad de conocer la posología de las dosis.

El algoritmo tiene un parámetro, Wt que puede ajustarse según el tipo de medicación estudiado, menor para fármacos de corta duración (como antibióticos) o mayor para tratamientos de larga duración (como antidepresivos).

Desde un punto de vista clínico, el algoritmo “smooth” presenta una ventaja significativa al estimar la adherencia a tratamientos combinados [65], [66]. Los pacientes que padecen enfermedades crónicas que requieren múltiples medicamentos pueden mostrar niveles variables de adherencia a los distintos fármacos de su régimen. Esta variabilidad podría traducirse en resultados de salud dispares y conducir a conclusiones erróneas si el enfoque metodológico considera que la adherencia a todos los medicamentos es uniforme. Además, estos pacientes a menudo enfrentan otras condiciones agudas que implican la incorporación de nuevos medicamentos o ajustes en las dosis, al mismo tiempo que mantienen su régimen habitual de medicación.

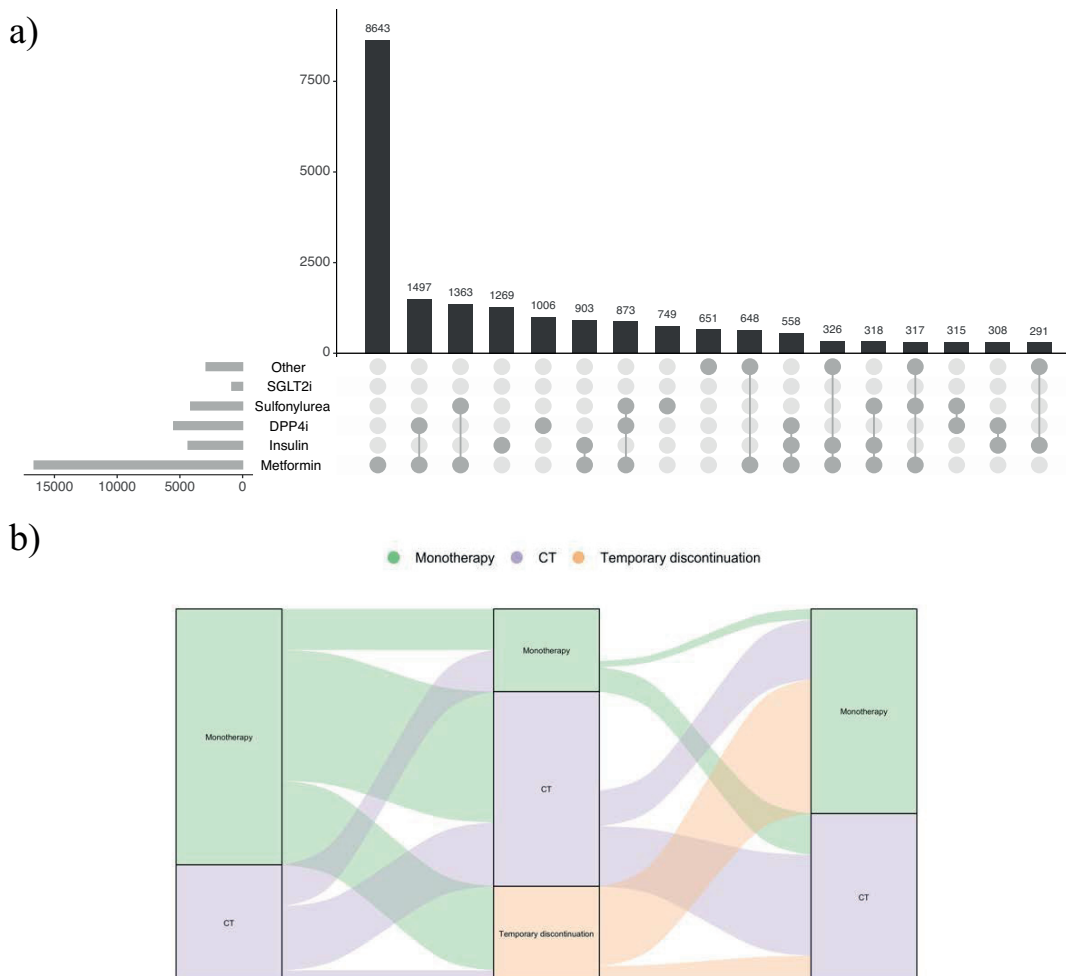
Otra área de investigación en la que nuestro algoritmo podría resultar útil es el estudio de eventos adversos debidos a la falta de efectividad de medicamentos, como los antibióticos o los tratamientos para la hipertensión, así como las interacciones entre medicamentos, como la combinación de anticoagulantes y antiinflamatorios no esteroides durante periodos breves.

El algoritmo “smooth” representa una solución automatizada para analizar datos de registros electrónicos de salud, garantiza coherencia metodológica, busca mejorar la calidad de los datos antes de su análisis y minimiza las discrepancias entre estudios que utilizan bases de datos similares proporcionando resultados coherentes y replicables en diferentes estudios. También evita realizar suposiciones antes de procesar los datos, lo que suele influir en los resultados de los análisis e introducir sesgos [64].

Visualización de los Tratamientos

Con el algoritmo también se facilita la capacidad de representar gráficamente los datos de exposición al tratamiento. Para explorar los patrones longitudinales de tratamiento, se utilizaron diversas técnicas visuales. Se emplearon gráficos de conjuntos (*upset chart*, Figura 20a) [67], [68] para representar la distribución de tratamientos solos o en combinación, en cada línea terapéutica, y diagrama de aluviales [69] para visualizar el flujo de cambios entre la monoterapia y la CT (Figura 20b).

Figura 20. Gráfico de Conjuntos (*Upset*) y Diagrama Aluvial



El panel a) es el gráfico de conjuntos con los tratamientos ADM de primar línea solos o combinados ordenados por frecuencia. En la parte inferior el diagrama aluvial representa el camino que hacen los pacientes con T2DM entre monoterapia y CT a lo largo del estudio.

Software para el Paquete Estadístico R

Está fuera de los objetivos de la tesis, pero con la idea de hacer el algoritmo accesible para todos los investigadores, se desarrolló un paquete de R que incluye todo el algoritmo con su manual de uso. Este paquete está disponible en el servidor oficial y para descarga de forma libre y gratuita (más detalle del paquete en el Apéndice 3).

6.2 Modelizar la Exposición a Medicamentos a partir del Algoritmo

Patrones Longitudinales de Tratamiento con Antidiabéticos

Para el segundo artículo de esta tesis, hemos llevado a cabo un análisis exhaustivo en un amplio grupo de pacientes incidentes con T2DM utilizando la base de datos SIDIAP. Nuestro objetivo principal fue describir el primer fármaco antidiabético prescrito, así como los patrones de tratamiento a lo largo del tiempo y su relación con el control glucémico medido mediante la HbA1c.

Para obtener estos patrones de tratamiento, modelamos las terapias de cada paciente utilizando el algoritmo “smooth”. Más detalles sobre este proceso se pueden encontrar en 4.2 *Diabetes Research and Clinical Practice*, que incluye el segundo artículo de la tesis.

Analizar Tratamientos Combinados

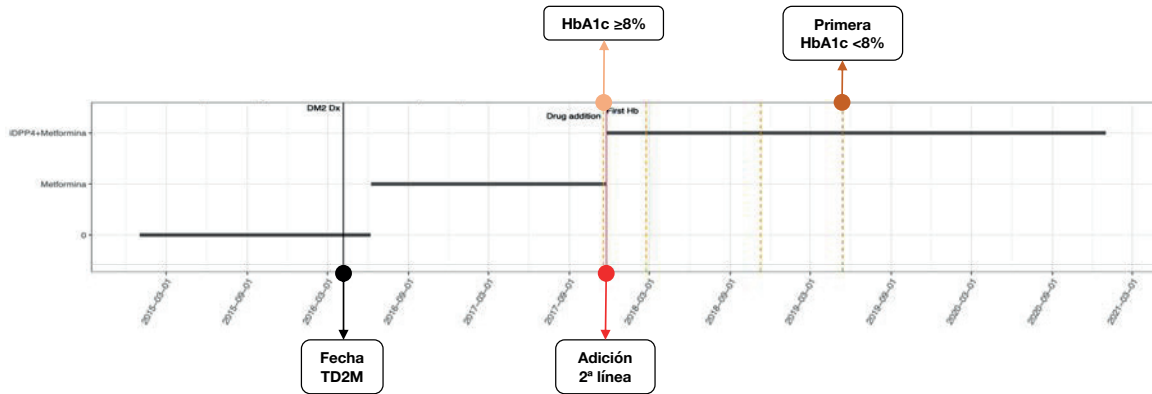
En el tercer y último artículo que conforma esta tesis, evaluamos el impacto de los tratamientos de segunda línea en los niveles de HbA1c y el peso en pacientes con T2DM que inician tratamiento con Metformina y no son capaces de controlar la enfermedad.

Selección de Pacientes

Durante la selección de la población, se modelaron los perfiles longitudinales paciente a paciente utilizando todos los registros de medicamentos antidiabéticos y usando el algoritmo “smooth”. Con especial interés en el momento que se introdujo el tratamiento de segunda línea y la duración del mismo, ya que la variable principal era el tiempo en días hasta el control de la HbA1c, una vez iniciada la terapia combinada. La Figura 21 proporciona un ejemplo de los diversos

eventos que se buscaron para seleccionar a los pacientes de estudio que cumplieran con los criterios de inclusión.

Figura 21. Perfil de un Paciente Analizado en el Tercer Artículo de Investigación



El gráfico muestra en el eje vertical el tratamiento al que está expuesto el paciente (0: no tratamiento). Las líneas verticales representan eventos de interés: en negro el momento del diagnóstico de T2DM y en rojo el momento de la adición de un tratamiento de segunda línea. Las mediciones de HbA1c realizadas al paciente se destacan con líneas discontinuas; en naranja cuando HbA1c $\geq 8\%$ y en marrón cuando la medición de HbA1c es $< 8\%$.

6.3 Limitaciones

En la Introducción (1.8 La Influencia de la Calidad de los Datos) ya se ha descrito el problema de obtener evidencias científicas sin una EHR de calidad. Algunas de estas limitaciones ya se conocían de antemano, sobre todo las que son de carácter teórico como los tipos de sesgo, formas de recogida de la información, problemas tecnológicos, etc. Sin embargo, durante el desarrollo de esta tesis, han surgido otras limitaciones más inesperadas que han dificultado el trabajo o nos ha condicionado a la hora de interpretar los resultados hallados.

Limitaciones del Algoritmo

Debido a la naturaleza del algoritmo de suavizar aquellos patrones complejos, es importante destacar que al usarlo podríamos perder exposiciones relevantes en cortos períodos de tiempo, así que “smooth” no es recomendable para todos los medicamentos. Por ejemplo, puede no ser eficaz en fármacos de acción prolongada.

Antes de utilizar el algoritmo, debemos establecer un valor para Wt , que nos permite escoger entre precisión y simplicidad. Este requisito es inherente al tipo de método estadístico que usa el algoritmo y puede ser un proceso complicado ya que seleccionar un valor adecuado requiere un conocimiento avanzado sobre los medicamentos estudiados y su comportamiento a largo plazo. Sin embargo, la simulación realizada nos permite fijar unos valores de referencia. En general, se sugiere usar un valor de Wt entre 30 y 60 días.

En la validación, hubiese sido preferible usar como comparador otro algoritmo de características similares. Sin embargo, en nuestra experiencia, no existen alternativas en la literatura (seguramente las haya en *software* de código cerrado o privado), y lo más común es encontrar métodos similares al tradicional con ligeras variaciones según el medicamento u objetivos del estudio. Por ejemplo, en proyectos realizados en colaboración con la Agencia Europea del Medicamento (EMA), nunca hemos encontrado un método o enfoque automatizado para tratar la exposición a medicamentos diferente al “tradicional” [70].

Desde la perspectiva de un científico de datos o analista, utilizar el algoritmo “smooth” es un proceso que consume tiempo y aplicar el algoritmo en grandes bases de datos como los EHR requiere una infraestructura informática en condiciones. El tiempo necesario para completar el proceso puede variar según el número de pacientes y el período de seguimiento, como referencia, en un ordenador moderno puede tardar entre 10 y 12 horas para procesar 1 millón de registros.

Finalmente, no es un algoritmo disponible en los programas estadísticos que hay actualmente en el mercado, lo que limita su uso en otros grupos de investigación. Por esta razón, para hacerlo más accesible, hemos desarrollado una librería para el paquete estadístico R¹², y que está disponible para descargar de forma gratuita en el repositorio oficial, el CRAN¹³.

¹² The R Project: <https://www.r-project.org>

¹³ The Comprehensive R Archive Network (CRAN): <https://cran.r-project.org>

Limitaciones en los Análisis de Exposición a Fármacos

Registros de Medicamentos

La más importante, es que en SIDIAP los registros de prescripciones electrónicas y de dispensaciones en farmacia no están vinculados entre sí. De hecho, existen situaciones en que una prescripción no tiene dispensación en farmacia o, lo que es más extraño, dispensaciones registradas que no tienen prescripción electrónica. Este último caso sucede sobre todo cuando el origen de la prescripción no se hace en un centro del ICS y, por lo tanto, no queda registrado en SISAP, fuente de origen de la base de datos SIDIAP. Para reducir este problema, filtramos por aquellas prescripciones cuyo origen fuera un centro del ICS o de Centros de Atención Primaria.

Además, no fue posible capturar información detallada sobre la posología en la base de datos SIDIAP, lo que nos impidió estimar la duración exacta del tratamiento a través de la dosis recetada. En consecuencia, nos vimos forzados a suponer que la dosis es constante durante todo el período de tratamiento. Sin embargo, esta suposición puede resultar algo restrictiva, dado que las dosis de los tratamientos pueden variar a lo largo del año como, por ejemplo, disminuir durante el verano o en situaciones de viaje, entre otros factores.

Finalmente, es importante destacar que carecemos de información acerca de si los cambios en el tratamiento fueron realizados por el mismo especialista o por diferentes profesionales de la salud, lo cual podría influir en los patrones de tratamiento observados. Tampoco tenemos acceso a datos que expliquen las razones detrás de los cambios de tratamiento debidos a eventos adversos, lo que nos impide explorar cualquier asociación entre los patrones de tratamiento a lo largo del tiempo y resultados como la mortalidad o hospitalizaciones.

Registros de Diagnósticos

Similar a lo que nos encontramos en los registros de medicamentos, en SIDIAP los diagnósticos no están vinculados a una prescripción electrónica. Sabemos que a los pacientes se les receta un medicamento pero no tenemos conocimiento exacto el problema de salud que se quiere tratar. Normalmente el diagnóstico y la prescripción se enlazan por proximidad de fechas, pero no siempre el resultado es preciso. Para no tener estos problemas, en estudios con SIDIAP se recomienda definir una población a partir del problema de salud, en nuestro caso el diagnóstico de T2DM y sobre dicha población de referencia se extraen el resto de datos, entre ellos los medicamentos prescritos. De nuevo, no es una aproximación infalible ya que siempre nos encontramos con un porcentaje de no tratados bastante elevado, del cual no sabemos con certeza la razón.

Adicionalmente, para los pacientes con T2DM, nos hemos encontrado en que la fecha de diagnóstico y el inicio del tratamiento es más cercana de lo esperado, lo que sugiere que el momento exacto del inicio de la enfermedad puede ser impreciso o no estar adecuadamente registrado. Aunque intentamos abordar esta cuestión limitando la selección de pacientes a aquellos con un registro de prescripción en un margen de 90 días antes o después de la fecha de diagnóstico, este enfoque también tiene sus limitaciones y no resuelve completamente el problema.

Limitaciones con el Tiempo de Seguimiento de los Pacientes

Es relevante mencionar que la desigualdad en los períodos de seguimiento entre los pacientes de nuestros estudios puede potencialmente afectar los resultados que observamos. Con el propósito de minimizar este efecto, aplicamos un criterio de seguimiento mínimo de al menos un año para la inclusión de los pacientes. Sin embargo, en pacientes con períodos de seguimiento relativamente breves, nuestra capacidad para modelar los tratamientos queda limitada, ya que el paciente carece de registros. Por el contrario, los pacientes con períodos de seguimiento más extensos se correlacionan con una mayor probabilidad de recibir más terapias, más eventos diagnósticos, efectos adversos, etc. Este fenómeno aumenta el sesgo sobre los análisis y los resultados obtenidos deben ser interpretados con cautela. Se requiere de investigaciones adicionales para abordar adecuadamente estas limitaciones.

Limitaciones de los Datos en Pacientes con T2DM

En particular, el grupo de pacientes tratados con GLP1 presenta una limitación importante ya que este tratamiento solo es reembolsado para pacientes con obesidad en el Servicio Español de Salud. Esto introduce un sesgo de selección, lo que significa que este grupo no es directamente comparable con otros grupos. En cuanto a los pacientes tratados con insulina, tampoco son fácilmente comparables ya que suelen ser aquellos con un control glucémico deficiente. Por lo tanto, elegimos el grupo de sulfonilureas como grupo de referencia para nuestras comparaciones estadísticas ya que es más comparable con el resto y clínicamente es considerado el tratamiento de segunda línea estándar [42].

Una limitación adicional de nuestros estudios con pacientes en T2DM radica en la falta de información detallada sobre eventos hipoglucémicos debido al registro deficiente en el sistema de atención primaria de Cataluña. Como las sulfonilureas son prevalentes en nuestra población de estudio hubiese sido interesante ajustar los análisis por estos eventos ya que las sulfonilureas se sabe que conllevan un riesgo de hipoglucemia. De nuevo, es importante interpretar los resultados teniendo esta limitación en mente.

7. Conclusiones

A continuación se listan las conclusiones de esta tesis doctoral.

1. La hipótesis general de la tesis “los estudios farmacológicos basados en EHR requieren el desarrollo de métodos que garanticen la obtención de evidencias científicas confiables y reproducibles” se trabajó mediante la creación de un novedoso algoritmo llamado “smooth”.
2. El “smooth” es un método automático que calcula las tendencias más probables de tratamientos en EHR. Estandariza, simplifica y mejora la fase de limpieza y manipulación de los datos antes de los análisis.
3. El ensayo de validación con cuatro expertas farmacólogas independientes demostró la capacidad del algoritmo para adaptarse a diferentes medicamentos y perfiles de tratamientos. El algoritmo es flexible y se puede usar en cualquier contexto, especialmente en estudios con pacientes que reciben múltiples medicamentos a lo largo de períodos prolongados.
4. El algoritmo facilita la modelación de la exposición a medicamentos usando EHR. Ayudó a detectar eventos de interés farmacológico, como las adiciones, deleciones, cambios en la terapia, inicio o finalización de terapias combinadas, discontinuaciones e interrupciones de tratamiento.
5. El algoritmo fue empleado en dos estudios que involucraron a pacientes con tratamiento para la diabetes tipo 2, de los cuales se derivaron las siguientes conclusiones:
 - a. El primer estudio abordó un análisis detallado del tratamiento antidiabético en pacientes recién diagnosticados con T2DM, desde la prescripción inicial hasta la tercera línea de tratamiento. Utilizando el “smooth”, se logró modelar con facilidad el momento de transición a la segunda y tercera línea, tanto en monoterapia como en terapia combinada, facilitando también la obtención de cálculos temporales para cada tratamiento.
 - b. El segundo estudio se propuso analizar los cambios en el peso y la HbA1c después de iniciar un tratamiento combinado de segunda línea en pacientes con un control deficiente de la enfermedad. El algoritmo “smooth” permitió detectar fácilmente el momento de cambio de la monoterapia a la terapia combinada de segunda línea, así como la duración de esta terapia hasta lograr el control de la enfermedad.

6. Con el “smooth” fue posible analizar el cumplimiento de forma longitudinal y entender su evolución a lo largo del tiempo. En ambos estudios se calculó el MPR para evaluar el cumplimiento, y los resultados obtenidos fueron precisos y se situaron dentro de los márgenes de cumplimiento esperado.
7. A partir de los datos suavizados por el algoritmo “smooth”, se describieron detalladamente los perfiles longitudinales de tratamiento en pacientes con T2DM. Estos perfiles permitieron analizar los patrones de tratamiento, los cambios más frecuentes entre la monoterapia y la terapia combinada, y su asociación con las fluctuaciones en la HbA1c.
8. De los resultados obtenidos en los estudios se obtuvieron las siguientes conclusiones principales:
 - a. Se observó que la pauta más común, desde la primera hasta la tercera línea, consistió en iniciar con metformina en monoterapia, añadir un segundo agente antidiabético en la segunda línea, y luego regresar a metformina en monoterapia.
 - b. El efecto de los tratamientos de segunda línea en el control de la HbA1c fue positivo en la mayoría de los pacientes, siendo las combinaciones de GLP1 e Insulina las que lograron un control más temprano en comparación con la combinación con Sulfonilurea.
 - c. La adherencia a la terapia demostró tener un efecto significativo en la reducción del tiempo necesario para alcanzar el control de la HbA1c en todas las combinaciones, excepto las que incluían SGLT2i.
 - d. Se observó que el tratamiento con GLP1 condujo a una reducción de peso más temprana y la combinación con SGLT2i fue efectiva en lograr reducciones significativas del peso corporal.

8. Nuevas Líneas de Investigación

Durante el desarrollo de esta tesis doctoral, surgieron nuevas líneas de investigación con el objetivo de mejorar el algoritmo presentado y llevar la modelización de la exposición a tratamientos a lo largo del tiempo y los análisis globales a un nivel superior.

En relación al algoritmo “smooth”, se han identificado tres aspectos con margen de mejora que pueden ser objeto de futuros desarrollos:

1. **Distribución de pesos:** actualmente, la distribución del peso que se asigna a cada tratamiento dentro de la ventana de suavizado W_t es uniforme, lo que significa que la probabilidad de asignar un tratamiento u otro es la misma para todos. Aunque esta simplificación es útil, desde un punto de vista clínico, podría no ser la mejor opción en todas las situaciones. Una alternativa sería repartir los pesos de manera exponencial (utilizando *Exponential Moving Averages*), lo que daría más peso a los tratamientos recientes en el tiempo. Además, se podría mejorar aún más la inteligencia del algoritmo al permitir que la asignación de pesos se base en las características específicas del medicamento en estudio o de la patología. Por ejemplo, se podría diseñar el algoritmo para que priorice automáticamente las adiciones o las terapias combinadas en función de la situación clínica.
2. **Mejoras en el No Tratamiento:** otro aspecto a trabajar más en el algoritmo “smooth” es el algoritmo cuando no hay tratamiento. Se relaciona con el punto anterior, ya que la ponderación debería ser diferente cuando lo más común dentro de la ventana de suavizado es la ausencia de tratamiento.
3. **Eficiencia de cálculo:** actualmente, “smooth” es completamente utilizable en estudios rutinarios, pero existe incertidumbre sobre su desempeño si se aplica en un *Common Data Model*. Además, siempre es posible mejorar la velocidad del algoritmo si se dedica tiempo y esfuerzo a optimizar su eficiencia.

Estamos seguros de que usar el “smooth” es recomendable en cualquier estudio con datos de medicamentos en EHR. Sin embargo, planeamos llevar a cabo más validaciones para respaldar aún más su valor:

- En primer lugar, tenemos la intención de comparar los tratamientos obtenidos por el algoritmo con la realidad. Esto se podría lograr utilizando registros más controlados de pacientes, como los que provienen de ensayos clínicos. Nuestro objetivo es analizar con precisión la capacidad del algoritmo para estimar el tratamiento real en lugar del tratamiento más probable.
- Otra validación de interés sería comparar los resultados finales de un estudio utilizando tanto el algoritmo "smooth" como otro método más "tradicional" de manera paralela. Al tener ambos conjuntos de resultados, podríamos investigar si hay diferencias en los cálculos de adherencia y cumplimiento, así como en los resultados finales, especialmente en medidas de asociación como las OR o las HR de las regresiones de Cox. Estudios previos han sugerido que estos estadísticos pueden variar significativamente en función del método utilizado durante la fase de limpieza de datos y definición de variables [71].
- Finalmente, nos ha quedado pendiente comparar el cumplimiento obtenido a partir de los datos del algoritmo contra el cumplimiento calculado a partir de los registros de la receta electrónica, más cercano a la realidad.

Estas futuras validaciones servirán para fortalecer la confianza en el algoritmo y proporcionar evidencia adicional de su utilidad en la investigación farmacológica.

La metodología utilizada para modelar los tratamientos a lo largo del tiempo es otro aspecto que requiere una mayor investigación. En la presente tesis, se explican los cálculos de diversos eventos, como adiciones, deleciones, interrupciones, cumplimiento, terapias simples y combinadas, así como cambios de tratamiento. Estas definiciones son bastante generales en la literatura, pero también podrían beneficiarse de algoritmos que determinen estos cálculos de manera automática, más precisa y generalizable entre estudios.

Además, es importante destacar que la relación entre el diagnóstico y la prescripción electrónica en SIDIAP se considera poco fiable. Aún siendo un caso un poco específico, necesitamos herramientas y mayor investigación sobre los datos que disponemos para así asegurar que obtenemos resultados fiables y de calidad.

9. Referencias

- [1] R. E. Sherman *et al.*, “Real-World Evidence — What Is It and What Can It Tell Us?,” *N. Engl. J. Med.*, vol. 375, no. 23, pp. 2293–2297, Dec. 2016, doi: 10.1056/NEJMs1609216.
- [2] C. Nabhan, A. Klink, and V. Prasad, “Real-world Evidence - What Does It Really Mean?,” *JAMA Oncology*, vol. 5, no. 6. 2019, doi: 10.1001/jamaoncol.2019.0450.
- [3] A. Makady, A. de Boer, H. Hillege, O. Klungel, and W. Goettsch, “What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews,” *Value Heal.*, vol. 20, no. 7, pp. 858–865, 2017, doi: 10.1016/j.jval.2017.03.008.
- [4] A. S. Motrinchuk, A. R. Kasimova, and E. A. Novoderezhkina, “NICE guidance to working with real-world data,” *Real-World Data Evid.*, vol. 2, no. 4, 2023, doi: 10.37489/2782-3784-myrd-25.
- [5] S. F. Garbade *et al.*, “Quantitative retrospective natural history modeling for orphan drug development,” *Journal of Inherited Metabolic Disease*, vol. 44, no. 1. 2021, doi: 10.1002/jimd.12304.
- [6] K. Häyrynen, K. Saranto, and P. Nykänen, “Definition, structure, content, use and impacts of electronic health records: A review of the research literature,” *Int. J. Med. Inform.*, vol. 77, no. 5, pp. 291–304, 2008, doi: 10.1016/j.ijmedinf.2007.09.001.
- [7] D. Blumenthal and M. Tavenner, “The ‘Meaningful Use’ Regulation for Electronic Health Records,” *N. Engl. J. Med.*, vol. 363, no. 6, 2010, doi: 10.1056/nejmp1006114.

- [8] J. A. Casey, B. S. Schwartz, W. F. Stewart, and N. E. Adler, “Using Electronic Health Records for Population Health Research: A Review of Methods and Applications,” *Annu. Rev. Public Health*, vol. 37, pp. 61–81, 2016, doi: 10.1146/annurev-publhealth-032315-021353.
- [9] J. R. Rogers, J. Lee, Z. Zhou, Y. K. Cheung, G. Hripcsak, and C. Weng, “Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review,” *J. Am. Med. Informatics Assoc.*, vol. 28, no. 1, pp. 144–154, 2021, doi: 10.1093/jamia/ocaa224.
- [10] L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal, “Electronic health records: Implications for drug discovery,” *Drug Discovery Today*. 2011, doi: 10.1016/j.drudis.2011.05.009.
- [11] S. E. Davis *et al.*, “Use of Electronic Health Record Data for Drug Safety Signal Identification: A Scoping Review,” *Drug Safety*. 2023, doi: 10.1007/s40264-023-01325-0.
- [12] S. D. Ramsey *et al.*, “Using electronic health record data to identify comparator populations for comparative effectiveness research,” *Journal of Medical Economics*, vol. 23, no. 12, pp. 1618–1622, 2020, doi: 10.1080/13696998.2020.1840113.
- [13] K. Yamamoto, S. Matsumoto, K. Yanagihara, S. Teramukai, and M. Fukushima, “A data-capture system for post-marketing surveillance of drugs that integrates with hospital electronic health records,” *Open Access J. Clin. Trials*, vol. 3, pp. 21–26, 2011, doi: 10.2147/OAJCT.S19579.
- [14] K. Park, “The use of real-world data in drug repurposing,” *Transl. Clin. Pharmacol.*, vol. 29, no. 3, pp. 117–124, 2021, doi: 10.12793/tcp.2021.29.e18.

- [15] M. Edouard, R. Hannah, R.-G. Lucas, A. Cameron, G. Charlie, and H. Joe, “Coronavirus Pandemic (COVID-19),” *Our World In Data*, 2020. <https://ourworldindata.org/coronavirus> (accessed Aug. 15, 2023).
- [16] C. M. Benedum *et al.*, “Replication of Real-World Evidence in Oncology Using Electronic Health Record Data Extracted by Machine Learning,” *Cancers (Basel)*., vol. 15, no. 6, 2023, doi: 10.3390/cancers15061853.
- [17] G. Sotorra-Figuerola, D. Ouchi, M. Giner-Soriano, and R. Morros, “Impact of adherence to drugs for secondary prevention on mortality and cardiovascular morbidity: A population-based cohort study. IMPACT study,” *Pharmacoepidemiol. Drug Saf.*, vol. 30, no. 9, pp. 1250–1257, 2021, doi: 10.1002/pds.5261.
- [18] H. M. Belli *et al.*, “A behavioral economics-electronic health record module to promote appropriate diabetes management in older adults: Protocol for a pragmatic cluster randomized controlled trial,” *JMIR Res. Protoc.*, vol. 10, no. 10, p. 1DUMMUY, 2021, doi: 10.2196/28723.
- [19] J. C. Denny *et al.*, “PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations,” *Bioinformatics*, vol. 26, no. 9, pp. 1205–1210, 2010, doi: 10.1093/bioinformatics/btq126.
- [20] E. Eldridge, E. Corbett, J. Jones, S. Mahmood, and N. D. Lin, “Characterizing ‘long-COVID’ using real world data: Post-discharge clinical course among patients initially hospitalized for COVID-19,” *Pharmacoepidemiol. Drug Saf.*, vol. 30, no. SUPPL 1, 2021.
- [21] A. Al-Aswad and S. Brownsell, “A Review Paper of the Current Status of Electronic Health Records Adoption Worldwide: The Gap between Developed and Developing

- Countries,” ... *Dev. Ctries.*, vol. 7, no. 2, 2013.
- [22] M. T. Woldemariam and W. Jimma, “Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: A systematic review,” *BMJ Heal. Care Informatics*, vol. 30, no. 1, pp. 8–10, 2023, doi: 10.1136/bmjhci-2022-100704.
- [23] L. M. Shapiro and R. N. Kamal, “Implementation of Electronic Health Records During Global Outreach: A Necessary Next Step in Measuring and Improving Quality of Care,” *Journal of Hand Surgery*, vol. 47, no. 3. pp. 279–283, 2022, doi: 10.1016/j.jhsa.2021.09.016.
- [24] Thiel R, Lupiáñez-Villanueva F, Deimel L, Gunderson L, and Sokolyanskaya A, *Interoperability of Electronic Health Records in the EU | Shaping Europe’s digital future*. 2021.
- [25] E. Medicines Agency, “A Common Data Model for Europe? - Why? Which? How? - workshop report,” vol. 44, no. October, 2018, [Online]. Available: www.ema.europa.eu/contact.
- [26] A. Frutos Pérez-Surio, F. Sala-Piñol, and N. Sanmartí-Martinez, “Evidence-based medicine clinical practice guidelines in the Spanish National Health System,” *Eur. J. Hosp. Pharm.*, vol. 23, no. 6, 2016, doi: 10.1136/ejhpharm-2015-000836.
- [27] D. Horgan *et al.*, “European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare,” *Healthc.*, vol. 10, no. 9, 2022, doi: 10.3390/healthcare10091629.
- [28] D. Elvira and P. Perez-Sust, “Pla Director de Sistemes d’Informació del SISCAT,” *Construint junts una estratègia de salut digital per a Catalunya*. 2017.

- [29] J. Darbà, “PARIS.Programa d’analítica de dades per a la recerca i la innovació en salut ,” *Agència Pública de Salut*, 2020. .
- [30] F. Fina, L. Méndez, and M. Medina, “Sistema de Información de los Servicios de Atención Primaria. La experiencia 2006-2008 del Institut Català de la Salut.,” *RISAI*, vol. 1, no. 1, 2008.
- [31] M. Recalde *et al.*, “Data Resource Profile: The Information System for Research in Primary Care (SIDIAP),” *Int. J. Epidemiol.*, vol. 51, no. 6, pp. e324–e336, Dec. 2022, doi: 10.1093/ije/dyac068.
- [32] B. Bolívar *et al.*, “Base de datos SIDIAP: la historia clínica informatizada de Atención Primaria como fuente de información para la investigación epidemiológica,” *Med. Clin. (Barc).*, vol. 138, no. 14, pp. 617–621, May 2012, doi: 10.1016/j.medcli.2012.01.020.
- [33] H. Støvring, A. Pottegård, and J. Hallas, “Determining prescription durations based on the parametric waiting time distribution,” *Pharmacoepidemiol. Drug Saf.*, vol. 25, no. 12, pp. 1451–1459, Dec. 2016, doi: 10.1002/pds.4114.
- [34] H. Lyu *et al.*, “Delayed denosumab injections and bone mineral density response: An Electronic Health Record-based Study,” *Journal of Clinical Endocrinology and Metabolism*. 2020, doi: 10.1210/clinem/dgz321.
- [35] J. Mascarenhas, M. Mehra, J. He, R. Potluri, and C. Loeffgren, “Patient characteristics and outcomes after ruxolitinib discontinuation in patients with myelofibrosis,” *J. Med. Econ.*, vol. 23, no. 7, pp. 721–727, 2020, doi: 10.1080/13696998.2020.1741381.
- [36] J. Sharman, S. M. Kabadi, J. Clark, and D. Andorsky, “Treatment patterns and

- outcomes among mantle cell lymphoma patients treated with ibrutinib in the United States: a retrospective electronic medical record database and chart review study,” *Br. J. Haematol.*, no. October 2020, pp. 737–746, 2020, doi: 10.1111/bjh.16922.
- [37] X. Li, S. R. Cole, D. Westreich, and M. A. Brookhart, “Primary non-adherence and the new-user design,” *Pharmacoepidemiol. Drug Saf.*, vol. 27, no. 4, pp. 361–364, 2018, doi: 10.1002/pds.4403.
- [38] M. A. Suchard *et al.*, “Comprehensive comparative effectiveness and safety of first-line antihypertensive drug class,” *Lancet*, vol. 394, no. 10211, pp. 1816–1826, 2019, doi: 10.1016/S0140-6736(19)32317-7.Comprehensive.
- [39] “2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022,” *Diabetes Care*, vol. 45, no. Supplement_1, pp. S17–S38, Jan. 2022, doi: 10.2337/dc22-S002.
- [40] E. Fau *et al.*, “Pautes per al tractament farmacològic de la diabetis mellitus tipus 2 Versió del 21 d’abril de 2021*,” *Scientia*, vol. 3, 2021.
- [41] J. Bian *et al.*, “Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data,” *J. Am. Med. Informatics Assoc.*, vol. 27, no. 12, pp. 1999–2010, 2020, doi: 10.1093/jamia/ocaa245.
- [42] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. 1, pp. 144–151, 2013, doi: 10.1136/amiajnl-2011-000681.
- [43] N. Menachemi and T. H. Collum, “Benefits and drawbacks of electronic health

- record systems,” *Risk Manag. Healthc. Policy*, vol. 4, pp. 47–55, 2011, doi: 10.2147/RMHP.S12985.
- [44] S. V. Wang *et al.*, “Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0,” *Pharmacoepidemiol. Drug Saf.*, 2017, doi: 10.1002/pds.4295.
- [45] C. Wang, Y. Gao, L. Zhu, M. Huang, Y. Wu, and J. Xuan, “Treatment Patterns in Patients With Newly Diagnosed Type 2 Diabetes in China: A Retrospective, Longitudinal Database Study,” *Clin. Ther.*, vol. 41, no. 8, pp. 1440–1452, 2019, doi: 10.1016/j.clinthera.2019.05.003.
- [46] M. Sharma, I. Nazareth, and I. Petersen, “Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: A retrospective cohort study,” *BMJ Open*, vol. 6, no. 1, 2016, doi: 10.1136/bmjopen-2015-010210.
- [47] G. Carney *et al.*, “Treatment pattern trends of medications for type 2 diabetes in British Columbia, Canada,” *BMJ Open Diabetes Res. Care*, vol. 10, no. 6, pp. 1–12, 2022, doi: 10.1136/bmjdr-2022-002995.
- [48] D. Care and S. S. Suppl, “6 . Glycemic Targets : Standards of Medical Care in Diabetes — 2022,” vol. 45, no. January, pp. 83–96, 2022.
- [49] E. Fau *et al.*, “Pautes per al tractament farmacològic de la diabetis mellitus tipus 2,” *Scientia*, 2013, Accessed: Jul. 14, 2022. [Online]. Available: <http://scientiasalut.gencat.cat/handle/11351/1403>.
- [50] J. J. Alemán *et al.*, *Guía de diabetes tipo 2 para clínicos: Recomendaciones de la redGDPS*. 2018.

- [51] O. Montvida, J. Shaw, J. J. Atherton, F. Stringer, and S. K. Paul, “Long-term trends in antidiabetes drug usage in the U.S.: Real-world evidence in patients newly diagnosed with type 2 diabetes,” *Diabetes Care*, vol. 41, no. 1, pp. 69–78, 2018, doi: 10.2337/dc17-1414.
- [52] M. Mata-Cases *et al.*, “Trends in the Degree of Control and Treatment of Cardiovascular Risk Factors in People With Type 2 Diabetes in a Primary Care Setting in Catalonia During 2007–2018,” *Front. Endocrinol. (Lausanne)*, vol. 12, no. January, pp. 1–11, 2022, doi: 10.3389/fendo.2021.810757.
- [53] O. J. Phung, D. M. Sobieraj, S. S. Engel, and S. N. Rajpathak, “Early combination therapy for the treatment of type 2 diabetes mellitus: Systematic review and meta-analysis,” *Diabetes, Obes. Metab.*, vol. 16, no. 5, pp. 410–417, 2014, doi: 10.1111/dom.12233.
- [54] N. A. Elsayed *et al.*, “9. Pharmacologic Approaches to Glycemic Treatment: Standards of Care in Diabetes—2023,” *Diabetes Care*, vol. 46, no. supp, pp. S140–S157, 2023, doi: 10.2337/dc23-S009.
- [55] A. Nicolucci *et al.*, “Treatment patterns and associated factors in 14 668 people with type 2 diabetes initiating a second-line therapy: Results from the global DISCOVER study programme,” *Diabetes, Obes. Metab.*, vol. 21, no. 11, pp. 2474–2485, 2019, doi: 10.1111/dom.13830.
- [56] M. J. Davies *et al.*, “Management of hyperglycaemia in type 2 diabetes, 2022. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD),” *Diabetologia*, vol. 65, no. 12, pp. 1925–1966, 2022, doi: 10.1007/s00125-022-05787-2.

- [57] M. Gil-Girbau *et al.*, “Reasons for medication non-initiation: A qualitative exploration of the patients’ perspective,” *Res. Soc. Adm. Pharm.*, vol. 16, no. 5, pp. 663–672, 2020, doi: 10.1016/j.sapharm.2019.08.002.
- [58] G. B. Sarbacker and E. M. Urteaga, “Adherence to insulin therapy,” *Diabetes Spectr.*, vol. 29, no. 3, 2016, doi: 10.2337/diaspect.29.3.166.
- [59] S. E. Andrade, K. H. Kahler, F. Frech, and K. A. Chan, “Methods for evaluation of medication adherence and persistence using automated databases,” *Pharmacoepidemiol. Drug Saf.*, vol. 15, no. 8, pp. 565–574, 2006, doi: 10.1002/pds.1230.
- [60] F. Pazan and M. Wehling, “Polypharmacy in older adults: a narrative review of definitions, epidemiology and consequences,” *Eur. Geriatr. Med.*, vol. 12, no. 3, pp. 443–452, 2021, doi: 10.1007/s41999-021-00479-3.
- [61] E. J. Hannan and J. Rissanen, “Recursive estimation of mixed autoregressive-moving average order,” *Biometrika*, vol. 70, no. 1, 1983, doi: 10.1093/biomet/70.1.303-b.
- [62] J. Durbin, “Efficient Estimation of Parameters in Moving-Average Models,” *Biometrika*, vol. 46, no. 3/4, p. 306, Dec. 1959, doi: 10.2307/2333528.
- [63] N. Sandgren, P. Stoica, and P. Babu, “On moving average parameter estimation,” in *European Signal Processing Conference*, 2012, pp. 2348–2351.
- [64] M. Żebrowska, P. Dzwiniel, and W. J. Waleszczyk, “Removal of the Sinusoidal Transorbital Alternating Current Stimulation Artifact From Simultaneous EEG Recordings: Effects of Simple Moving Average Parameters,” *Front. Neurosci.*, vol. 14, Jul. 2020, doi: 10.3389/fnins.2020.00735.

- [65] B. Vlachos *et al.*, “Analysis of the Adherence and Safety of Second Oral Glucose-Lowering Therapy in Routine Practice From the Mediterranean Area: A Retrospective Cohort Study,” *Front. Endocrinol. (Lausanne)*, vol. 12, no. July, pp. 1–12, 2021, doi: 10.3389/fendo.2021.708372.
- [66] J. M. Franklin *et al.*, “The relative benefits of claims and electronic health record data for predicting medication adherence trajectory,” *Am. Heart J.*, vol. 197, pp. 153–162, 2018, doi: 10.1016/j.ahj.2017.09.019.
- [67] A. Lex, N. Gehlenborg, H. Strobel, R. Vuilleumot, and H. Pfister, “UpSet: Visualization of intersecting sets,” *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1983–1992, 2014, doi: 10.1109/TVCG.2014.2346248.
- [68] J. R. Conway, A. Lex, and N. Gehlenborg, “UpSetR: An R package for the visualization of intersecting sets and their properties,” *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, 2017, doi: 10.1093/bioinformatics/btx364.
- [69] M. Rosvall and C. T. Bergstrom, “Mapping change in large networks,” *PLoS One*, vol. 5, no. 1, 2010, doi: 10.1371/journal.pone.0008694.
- [70] J. Cid-Ruzafa *et al.*, “Linaclotide utilization and potential for off-label use and misuse in three European countries,” *Therap. Adv. Gastroenterol.*, vol. 15, no. X, pp. 1–14, 2022, doi: 10.1177/17562848221100946.
- [71] S. R. Pye *et al.*, “Assumptions made when preparing drug exposure data for analysis have an impact on results: An unreported step in pharmacoepidemiology studies,” *Pharmacoepidemiol. Drug Saf.*, 2018, doi: 10.1002/pds.4440.
- [72] H. Xu *et al.*, “Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality,” *J. Am. Med.*

Informatics Assoc., vol. 22, no. 1, pp. 179–191, 2015, doi: 10.1136/amiajnl-2014-002649.

- [73] J. J. Hughey and J. M. Colby, “Discovering cross-reactivity in urine drug screening immunoassays through large-scale analysis of electronic health records,” *Clin. Chem.*, vol. 65, no. 12, pp. 1522–1531, 2019, doi: 10.1373/clinchem.2019.305409.
- [74] D. M. Bates and J. M. Chambers, “Programming with Data: A Guide to the S Language,” *Technometrics*, vol. 41, no. 3, 1999, doi: 10.2307/1270578.

10. Apéndices

Apéndice 1. EHR en SIDIAP

Tablas y Datos Disponibles en SIDIAP

La Tabla Suplementaria 1 contiene las principales datos y dominios disponibles en SIDIAP¹⁴. En resumen, a junio de 2021, contiene 8.036.948 de personas, >171 millones de diagnósticos y >1000 millones de prescripciones electrónicas o dispensaciones en farmacia.

Tabla Sup. 1. Descripción de las Tablas y Variables Disponibles en SIDIAP

Type Of Data Collected	Data Domain In The Sidiap Database	Number Of People With At Least One Entry For The Data Domain	Total Number Of Entries For The Data Domain	Key Information Recorded In The Data Domain
Socio-demographics	Population	8036948	8036948	Date of birth, sex, date of entry to the database, date and cause of exit (death, transferral out of the database) if applicable, nationality (grouped into 11 categories)
	Socio-economic variables	8036948	8036948	Individual income level (<18000€, between 18000€ and 100000€, >100000€ per year), deprivation indices measured at the census tract level (MEDEA, IP2011) and the basic health area level (ISC), requested pharmaceutical contribution (exempted, or contribution of 10%, 40%, 50%, 60%), maximum monthly pharmaceutical

¹⁴ Fuente: Tabla 2 del artículo publicado por Martina Recalde *et al.* en *International Journal of Epidemiology* en Abril de 2022. <https://doi.org/10.1093/ije/dyac068>

contribution (without limit, or up to 8.23€, 18.52€ or 61.75€)

	Regional variables	8036948	18036948	Type of residential area (rural or urban), province (Barcelona, Tarragona, Girona, Lleida) and basic health area (n = 388) where the person resides; productive unit (333 categories, e.g. EAP Vallirana, EAP Salt, EAP La Garriga, etc.), sanitary scope (10 categories, e.g. Barcelona, Girona, Terres del Ebre, etc.).
	Complexity	8036948	8036948	Clinical risk group (indicator based on the person's co-morbidities), state of fragility (categorizes persons into 'with chronic complexity', 'with advanced chronic disease' or none) and whether the person lives in a nursing home (yes, no)
Health conditions	Primary care diagnoses	7531524	171215580	ICD-10 code, dates of start and end (if there is any), SIDIAP grouper (e.g. 'Arthrosis' includes two ICD-10 codes with its descendants) and type of productive unit (hospital, primary care team, mental health centre, etc.) that registers the health problem
	Sick, maternity or paternity leaves	3191357	16651062	Coding of the health problem causing the sick leave (ICD-10) and dates of start and end of leave
Medications and vaccines	Prescriptions	7083845	320380795	ATC code, treatment group (e.g. anxiety, hormonal replacement therapy, hypertension), dates of start and end of prescription, posology, frequency of intake and DDD of the medication, speciality of the health professional (e.g. general medicine, gynaecology, paediatrician) and setting (hospital, primary care team, mental health centre, etc.) of prescription
	Dispensations	6928471	1082492571	ATC code, dates of start and end of the dispensation, DDD, number of packages dispensed per month
	Adverse reactions	240047	316214	ATC code of the drug that produces the adverse reaction and date of occurrence

	Vaccinations	6443204	116615695	Code, date and dose of administration and grouper of the antigen of the vaccine
Laboratory tests	Analytical variables	5703343	1129564883	Biomarker/measurement (e.g. glucose, cholesterol, bilirubin, PCR), date of test/measurement, result value and unit of the test, and interpretation of the value (in four categories: positive, negative, indeterminate, inconclusive or through free text if applicable)
	Serology	2051806	11170439	Serological test (e.g. hepatitis C antibodies, HIV) date and result (i.e. positive, negative, indeterminate)
Clinical practice and lifestyle information	Clinical and lifestyle variables	6545968	513637096	Clinical measurement (e.g. systolic/diastolic blood pressure, weight, height) and lifestyle variables (e.g. alcohol consumption, smoking status) date and value of the registry
	Request of complementary explorations/tests	4314375	21741943	Type (e.g. mammography, colonoscopy) and date of the exploration request
Visits	Visits	7473119	695049273	Visited service, SIDIAP grouper (e.g. dermatology, nursing, general practice), type of productive unit (hospital, primary care team, mental health centre, etc.) that does the visit, place (online, in person at the health centre or at the residence of the person) where the visit takes place
	Referrals	4738748	22042118	Health service to which the person is being referred to, date and cause (usually suspicion of an ICD-10 code) of referral
Sexual and reproductive health	Variables of sexual and reproductive health assistance	1566025	172218876	Variable (e.g. year of menopause, breastfeeding, or result of the last mammography), date and value (if none, the result is reported in another variable containing text in free format) of the registry

	Pregnancy	427193	649630	Date of last period and of estimated delivery, type of delivery (e.g. voluntary miscarriage, natural miscarriage, c-section), gestational age (in weeks), miscarriage risk (none, low, normal, moderate, high, very high), number of fetuses, trimestral obstetric ultrasounds [containing information about the mother (type of exploration, location of the placenta, type of amniotic fluid, etc.) and the fetus(es) (sex, cardiac activity, cephalic circumference, etc.)]
Paediatric health (<15 years of age)	Paediatric health	1592446	263953678	Variables related to birth (e.g. gestational age at delivery, sex, weight, height), screening (e.g. neonatal deafness, cystic fibrosis, congenital hypothyroidism), development (e.g. reaction to external stimuli, object manipulation, sphincter control), nutrition (e.g. gluten intolerance, breastfeeding length, beikost consumption), school (e.g. course, integration, performance), hygiene (e.g. correct fingernail, oral and genital hygiene), leisure and sports (e.g. extracurricular and weekend activities, screen time) and sleeping patterns (e.g. dyssomnia, sleeping schedule, snoring)
Other data sources available through external linkage	Diagnoses and medical procedures at hospital dischargee	3223390	59748073	ICD-10-CM/PCS code, SIDIAP grouper, position of the diagnosis/procedure at the time of admission (used to prioritize the diagnoses/procedures that caused the admission), dates and circumstance (urgent, scheduled) of admission, and dates and circumstance (eight categories, including home, voluntary discharge or death) of discharge. From 2016 onwards, type of anaesthesia used, ICU admission and length of it, date of the first ICU admission
	Hospital medication for outpatient dispensing	516557	14563331	ATC code, date of dispensation, content per box (tablets, syringes, suppositories, etc.), code of the pharmaceutical product, reason for the dispensation, price of the dispensation

Pacientes con T2DM en SIDIAP

Los pacientes con T2DM se identifican en función de los códigos diagnóstico ICD-10 listados en Tabla Sup. 2.

De los más de 8 millones en SIDIAP, 373,185 presentan diagnóstico de T2DM (datos hasta 2017¹⁵). La edad promedio es de 70.3 (SD=12.1) años, 37.1% mayores de 75 años, el 45.1% de los pacientes son mujeres, en media, tienen 8.7 (SD=6) años de seguimiento y el 44.9% són obesos. Más detalles de las características clínicas de esta población en la Tabla Sup. 3.

Tabla Sup. 2. Códigos ICD-10 que Identifican a los Pacientes con T2DM

DESCRIPTION	CIM-10
Type 2 diabetes mellitus without complications	E11.9
Type 2 diabetes mellitus with unspecified diabetic retinopathy without macular edema	E11.319
Type 2 diabetes mellitus with diabetic cataract	E11.36
Type 2 diabetes mellitus with diabetic polyneuropathy	E11.42
Type 2 diabetes mellitus with other diabetic kidney complication	E11.21
Type 2 diabetes mellitus with other diabetic kidney complication	E11.29
Type 2 diabetes mellitus with diabetic peripheral angiopathy without gangrene	E11.51
Type 2 diabetes mellitus with other specified complication	E11.69
Type 2 diabetes mellitus with other diabetic ophthalmic complication	E11.39
Type 2 diabetes mellitus with hyperglycemia	E11.65

¹⁵ Fuente: Tabla 1 de M.Mata *et al.* publicado en *Primary Care Diabetes* en Febrero de 2021. <https://doi.org/10.1016/j.pcd.2021.02.003>

Type 2 diabetes mellitus with unspecified complications	E11.8
Type 2 diabetes mellitus with other diabetic neurological complication	E11.49
Type 2 diabetes mellitus with other skin complications	E11.638
Type 2 diabetes mellitus with foot ulcer	E11.621
Type 2 diabetes mellitus	E11
Type 2 diabetes mellitus with other circulatory complications	E11.59
Type 2 diabetes mellitus with diabetic neuropathy, unspecified	E11.40
Type 2 diabetes mellitus with hyperosmolarity with coma	E11.01
Type 2 diabetes mellitus with unspecified diabetic retinopathy with macular edema	E11.311
Type 2 diabetes mellitus with diabetic autonomic (poly)neuropathy	E11.43
Type 2 diabetes mellitus with hypoglycemia with coma	E11.641
Type 2 diabetes mellitus with hypoglycemia without coma	E11.649
Type 2 diabetes mellitus with hyperosmolarity without nonketotic hyperglycemic-hyperosmolar coma (NKHHC)	E11.00
Type 2 diabetes mellitus with diabetic mononeuropathy	E11.41
Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema left eye	E11.3292
Type 2 diabetes mellitus with other skin ulcer	E11.622

Table Sup. 3. Características de los Pacientes con T2DM en SIDIAP

VARIABLES*	ALL PATIENTS**	OBESITY (IMC > 30	
		KG/M ²)	AGE ≥ 75 YEARS
N (%)	N = 373,185 (100%)	N = 144,592 (44.9%)	N = 138,374 (37.1%)
Mean age, years (SD)	70.3 (12.1)	68.1 (11.7)	82.6 (11.2)
Gender (female) %	45.1	49.1	55.6
Mean T2DM duration, years (SD)	8.7 (6.0)	8.8 (6.0)	11.2 (6.8)
Non-smoker, %	54.8	55.0	69.2
Current smoker, %	14.3	12.8	4.8
Former smoker, %	30.9	32.2	26.0
Mean Systolic blood pressure, mmHg (SD) (N = 353,331)	133.0 (13.6)	133.8 (13.3)	133.7 (14.4)
Mean Diastolic blood pressure, mmHg (SD) (N = 353,331)	75.0 (9.7)	76.5 (9.5)	71.4 (9.6)
Mean BMI, kg/m ² (SD) (N = 321,739)	30.1 (5.2)	34.5 (4.2)	28.9 (4.7)
<i>Comorbidities</i>			
Obesity (BMI > 30 kg/m ²), % (N = 321,739)	44.9	100	36.1
Hypertension, %	71.9	77.7	84.4
Hyperlipidaemia, %	61.1	61.8	60.4

Cardiovascular disease, %	23.2		21 32.3
Heart Failure, %	6.9	8.2	13.0
Retinopathy, %	13.0	13.2	15.8
Neuropathy, %	10.4	12.5	10.0
Chronic Renal Failure (eGFR < 60 ml/min/1.73 m ²), % (N = 336,198)	28.0	26.5	50.3
Albuminuria (UACR > 30 mg/dl), %	14.2	15.5	16.1
<i>Laboratory results</i>			
Mean HbA1c, % (SD) (N = 330,014)	7.1 (1.3)	7.3 (1.3)	7.0 (1.1)
HbA1c < 6.5%, %	34.7	31.0	37.7
HbA1c 6.5–6.9%, %	20.7	20.2	21.5
HbA1c 7–8.9%, %	35.5	39.4	35.5
HbA1c ≥ 9%, %	9.0	9.3	5.3
Mean total cholesterol, mg/dL (SD) (N = 335,522)	182.0 (39.9)	186.0 (36.1)	178.7 (35.4)
Mean LDL-c, mg/dL (SD) (N = 315,578)	103.0 (32.5)	104.9 (30.2)	100.6 (29.5)
Mean HDL-c, mg/dL (SD) (N = 315,578)	48.8 (12.9)	47.6 (11.7)	50.4 (12.7)

Mean Triglycerides, mg/dL (SD) (N = 328,291))	159.0 (107.0)	174.6 (104.3)	142.7 (69.3)
Mean eGFR, ml/min/1.73 m ² (SD) (N = 336,198)	76.8 (21.7)	78.2 (21.0)	63.7 (18.2)
<i>Antidiabetic treatment</i>			
Only lifestyle modification, N (%)	18.4	15.3	19.4
Non-Insulin antidiabetic drug Monotherapy, %	37.1	37.0	37.9
Non-Insulin antidiabetic drug Combination, %	23.2	24.1	19.7
Insulin monotherapy, %	6.0	5.1	8.5
Insulin + Non-Insulin antidiabetic drug, %	15.3	18.6	14.5
Metformin, %	66.3	70.1	58.4
Sulfonylureas, %	19.0	19.8	17.8
Repaglinide, %	4.9	4.9	6.7
DPP4i, %	17.0	17.8	16.1
SGLT2i, %	2.6	4.0	0.8
GLP1ra, %	1.4	2.9	0.3
Pioglitazone, %	0.8	1.1	0.6
Insulin, %	21.3	23.7	23.0

Apéndice 2. Validación del Algoritmo

En este apéndice se recopilan resultados adicionales obtenidos durante la validación para complementar aquellos presentados en el primer artículo de la tesis.

Validación Cualitativa

Con la colaboración de cuatro profesionales independientes, expertas en bases de datos y en estudios farmacoepidemiológicos con EHR, se llevó a cabo una validación cualitativa del algoritmo “smooth” en comparación tanto a los perfiles de tratamientos originales, como con respecto a otro método denominado “tradicional”. Cada especialista revisó un conjunto de 100 pacientes (Tabla Sup. 4), 25 pacientes fueron validados por todas las expertas y así analizar la consistencia de criterios entre las profesionales. Con el propósito de eliminar posibles sesgos, las revisoras trabajaron de manera ciega, sin conocer qué método o algoritmo se encontraba detrás de cada resultado.

El método tradicional es un enfoque más intuitivo y sencillo, comúnmente observado en la literatura, que se utiliza para reducir el ruido y la variabilidad en los registros electrónicos de medicamentos [43],[44]. En resumen, comienza con el primer tratamiento observado y acepta un cambio de tratamiento solo si el nuevo es más largo que un cierto período de tiempo. Este período es generalmente arbitrario, una suposición hecha por el investigador basada en las características del fármaco. Para nuestro estudio de validación, establecimos el período en 60 días, excepto para los antibióticos (el tratamiento a corto plazo), con un período de 15 días. En el caso del algoritmo “smooth”, se configuró el parámetro Wt en 60 días en los 4 escenarios.

Tabla Sup. 4. Descripción de los Grupos de Medicamentos Analizados en la Validación

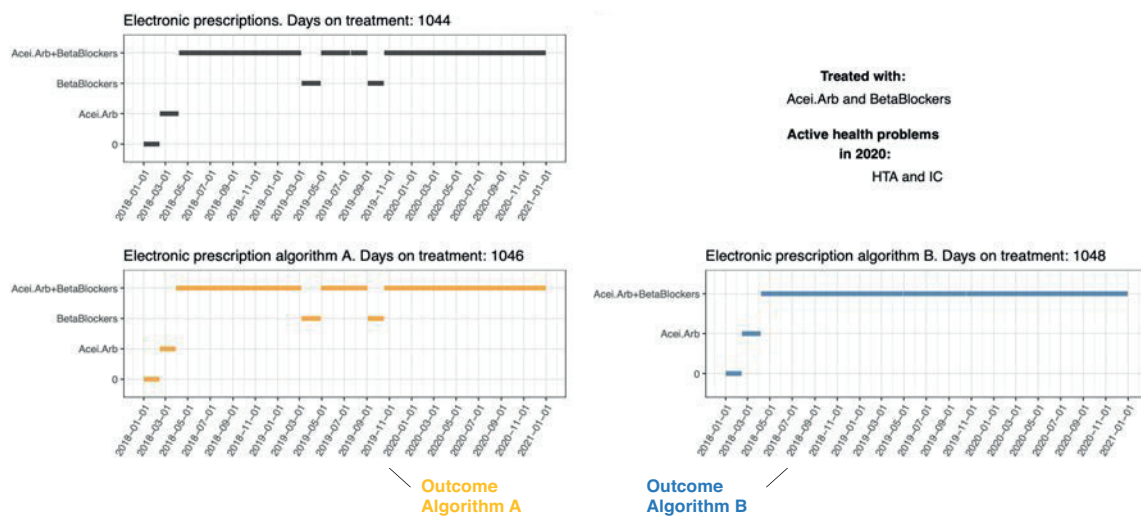
TIPO DE TRATAMIENTO	CÓDIGO ATC	DESCRIPCIÓN	#PRESCRIPCIONES EN LA BASE DE DATOS	#MUESTRAS ANALIZADAS	#MUESTRAS ANALIZADAS POR TODAS LAS REVISORAS
Combinación de tres medicamentos en la prevención cardiovascular	B01AC*	Antiplaquetarios	21,253,742	120 (30 por revisora)	60 (15 por revisora)
	C07*	Beta bloqueantes			
	C09A*, C09B*, C09C*, C09D*	IECA + ARA-II			
Tratamientos agudos a largo plazo con propensión a la interrupción	N06A*	Antidepresivos	3,859,496	80 (20 por revisora)	40 (10 por revisora)
Combinación a largo plazo de dos grupos farmacológicos para el tratamiento crónico de la T2DM	A10*	Insulinas y Antidiabéticos Orales	22,271,154	120 (30 por revisora)	60 (15 por revisora)
Tratamiento de corta duración	J01*	Antibióticos	10,846,282	80 (20 por revisora)	40 (10 por revisora)

En inglés.

En la Figura Sup 1, se presenta el ejemplo de un paciente revisado que está expuesto a tratamiento combinado de IECA, ARA-II, beta bloqueantes y antiplaquetarios (escenario de tratamiento combinado con tres medicamentos) para tratar la hipertensión. El panel superior izquierdo (en negro) muestra el gráfico original del perfil de exposición al tratamiento de anticoagulantes. En la parte inferior, se presentan los resultados después de aplicar el algoritmo “tradicional” o el “smooth” (etiquetados como A o B para mantener el ciego). En el lado izquierdo (en amarillo), se encuentra el resultado después de aplicar el algoritmo A. Finalmente, en la parte inferior derecha (en azul), se muestra el gráfico obtenido al aplicar el algoritmo B.

Para los 400 pacientes, se creó una figura del mismo estilo. A partir de estas, las expertas rellenaron una tabla (ejemplo en Tabla Sup. 5) en la que registraban la evaluación cualitativa. Para cada ítem, debían indicar si los resultados obtenidos a través de los algoritmos A y B mejoraron, empeoraron o no modificaron los perfiles originales de tratamiento. En primer lugar, cada revisora realizaba una validación global del resultado obtenido por cada algoritmo. Acto seguido, siempre que fuera posible, se evaluaba como los algoritmos A y B por separado ayudaban o no a detectar las discontinuaciones, los cambios y las suspensiones de tratamiento. Finalmente, de cada ítem, las revisoras podían comparar entre sí los dos algoritmos.

Figura Sup. 1. Ejemplo de un Paciente Analizado en el Estudio de Validación



La figura muestra un paciente expuesto a tratamiento combinado de ACEI, ARA-II y beta bloqueantes. Arriba a la izquierda se ilustra el gráfico original. A izquierda (en amarillo) el resultado de aplicar el algoritmo A y a derecha (en azul) el resultado de usar el algoritmo B.

Tabla Sup. 5. Ejemplo de Tabla de Validación Completada por Cada Revisora

Identificador	Aspecto a Validar	Comparación		
		Algoritmo A	Algoritmo B	A contra B
ejemplo fig. 14	Global respecto al Original	Igual	Mejor	B
ejemplo fig. 14	Discontinuación del tratamiento	No aplica	No aplica	-
ejemplo fig. 14	Cambios tratamiento	Igual	Mejor	B
ejemplo fig. 14	Parón tratamiento	Igual	Igual	-

**Validación de ejemplo, no resultado real.*

Antes de iniciar la validación de los 400 pacientes, se efectuó una sesión de formación con las cuatro revisoras. Esta sesión abarcó una introducción a los datos, los medicamentos

seleccionados y la explicación de los algoritmos. Además, sirvió para comentar, discutir y unificar los criterios que debían seguir durante la validación.

De los 400 pacientes revisados, el algoritmo no realizó modificaciones en los tratamientos de un 39% de los casos, y en el 56.8% de los pacientes contribuyó a mejorar dichos tratamientos. Estos resultados se mantuvieron consistentes en los diferentes grupos de tratamientos analizados. El algoritmo se diseñó específicamente para terapias combinadas (CT) y la validación en el grupo de tres o más medicamentos combinados (ACEI, ARA-II, Beta bloqueantes y Antiplaquetarios) obtuvo las calificaciones más favorables para el algoritmo "smooth". En esta categoría, solo empeoró el tratamiento en un 1.7% de los pacientes revisados, mientras que mejoró el tratamiento en el 69.2% restante (Tabla Sup. 6).

Tabla Sup. 6. Resultados de la Validación en los Diferentes Grupos de Medicamento

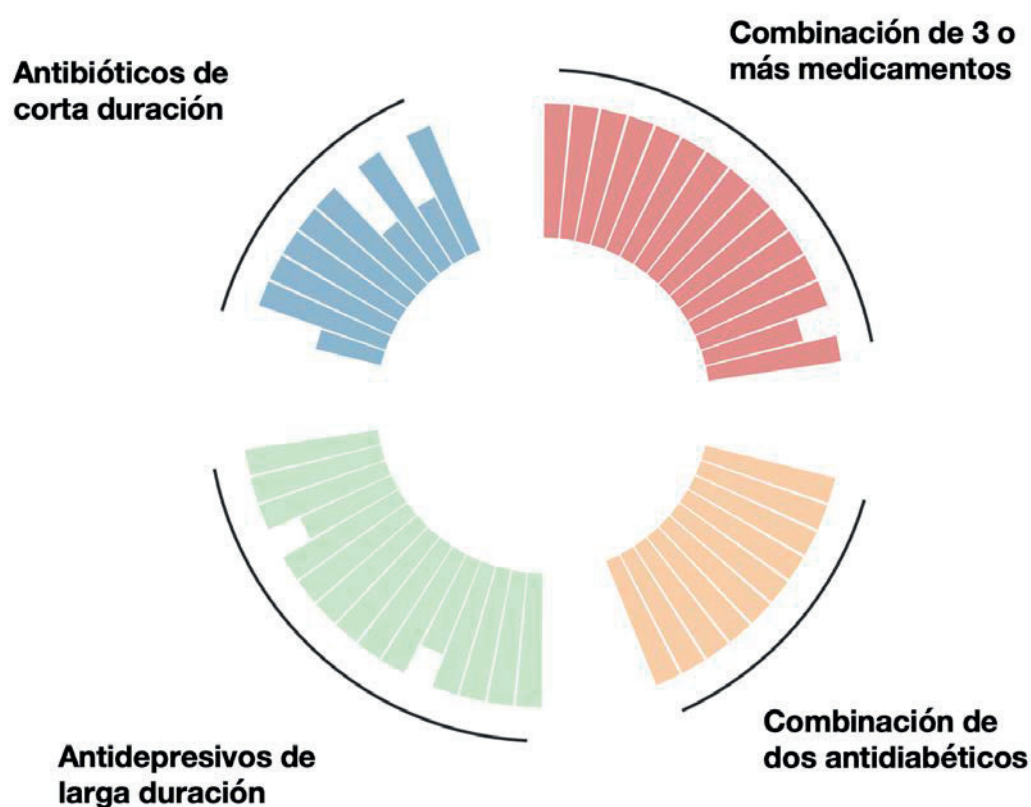
	GLOBAL	COMBINACIÓN DE 3 O MÁS MEDICAMENTOS	ANTIDEPRESIVOS DE LARGA DURACIÓN	COMBINACIÓN DE DOS FÁRMACOS ANTIDIABÉTICOS	ANTIBIÓTICOS SISTÉMICOS DE CORTA DURACIÓN
Algorithm "smooth" compared to the original data (n, %)					
N	400	120	80	120	80
<i>Peor</i>	17 (4.2)	2 (1.7)	1 (1.2)	6 (5.0)	8 (10.0)
<i>Igual</i>	156 (39.0)	35 (29.2)	30 (37.5)	42 (35.0)	49 (61.3)
<i>Mejor</i>	227 (56.8)	83 (69.2)	49 (61.3)	72 (60.0)	23 (28.7)

Un objetivo secundario que se planteó durante la validación fue el de comparar "smooth" con otro algoritmo "tradicional" utilizado en diferentes estudios [72], [73]. Aunque el algoritmo "tradicional" se consideró una buena aproximación, se observaron diferencias estadísticamente significativas entre ambos algoritmos. El algoritmo "tradicional" empeoró los tratamientos en un 18.5% de los pacientes y su rendimiento no fue uniforme entre los diferentes grupos de tratamientos. En particular, mostró un rendimiento subóptimo en el caso de los antibióticos sistémicos, pero se mostró bastante similar al algoritmo "smooth" en terapias combinadas de tres medicamentos (datos no mostrados). En resumen, las valoraciones realizadas por las profesionales fueron más positivas en el algoritmo "smooth".

Concordancia entre Revisoras

Cada revisora analizó 100 pacientes, sin embargo, 25 de estos eran repetidos y las cuatro revisoras los evaluaron. Así, con estos 25 se pudo hacer un análisis de “*agreement*” y valorar si el resultado en general concordaba. La Figura Sup. 2 contiene los resultados para cada paciente analizado. Cada barra representa a un paciente y la altura es el % de veces que las revisoras dieron el mismo veredicto.

Figura Sup. 2. Concordancia entre las Revisoras Durante la Validación



La figura ilustra los 25 pacientes (barras) coloreadas según el tipo de fármaco al que estaban expuestos. La altura de la barra representa el % de concordancia.

Apéndice 3. Paquete de R *smoothy*

R es un lenguaje de programación y un entorno de software diseñado especialmente para satisfacer las necesidades de aquellos que se dedican a la investigación y el análisis de datos. Su historia se remonta a principios de los años 90 cuando dos estadísticos, Ross Ihaka y Robert Gentleman, de la Universidad de Auckland en Nueva Zelanda, comenzaron a trabajar en el desarrollo de R como un proyecto académico. Se basaron en el lenguaje S [74], que era un lenguaje de programación desarrollado en los laboratorios Bell para análisis de datos y gráficos estadísticos.

El Repositorio CRAN

Un repositorio, en el contexto de la informática y la tecnología de la información, es un lugar donde se almacenan y gestionan datos, archivos, código fuente u otros recursos digitales de forma organizada y supervisada. Los repositorios son fundamentales para el control de versiones, la colaboración, el seguimiento de cambios y la distribución.

Dentro del ecosistema de R, el Comprehensive R Archive Network (CRAN) desempeña un papel fundamental. CRAN es un repositorio online que alberga numerosos paquetes y extensiones desarrollados por la comunidad de usuarios de R. Estos paquetes proporcionan una amplia gama de funcionalidades adicionales que los usuarios de R pueden aprovechar para realizar tareas específicas, desde análisis estadísticos avanzados hasta visualización de datos complejos, y CRAN funciona como plataforma de distribución.

Contribuir a CRAN:

Para compartir nuestro propio paquete de R con la comunidad a través de CRAN, es necesario seguir una serie de pasos en un proceso de revisión no muy diferente al de los artículos científicos. Concretamente, para incorporar el paquete *smoothy* en CRAN, pasamos por las siguientes etapas:

- **Desarrollo del paquete:** comenzamos desarrollando el paquete en un entorno de programación local. Para el desarrollo, seguimos la metodología detallada en la publicación al pie de la letra, ya que en CRAN se valora especialmente que cualquier nueva funcionalidad esté publicada en revista científica revisada por pares. Para pasar los criterios, hay que asegurarse de que el paquete sea útil, esté bien documentado y siga los estándares de calidad de CRAN.

- **Pruebas y verificaciones:** realizamos pruebas exhaustivas para asegurar que el paquete funciona sin errores. Verificamos que se podía usar en diferentes versiones de R y sistemas operativos como MacOS, Windows y Linux.
- **Documentación:** preparamos una documentación completa con ejemplos, tutoriales de uso y una lista de las funciones proporcionadas por el paquete (más información en el **Error! Not a valid bookmark self-reference.**).
- **Presentación a CRAN:** preparamos un archivo de presentación que incluya detalles sobre el paquete, autores y su ORCID, publicaciones de referencia para el paquete con DOI, licencia de uso (Licencia Pública General de GNU (GPLv3¹⁶)) y datos de contacto del administrador y persona que mantiene el paquete.
- **Envío a CRAN:** enviamos el paquete por internet al equipo de CRAN siguiendo sus pautas específicas de presentación de paquetes. Esto incluye pasar una serie de controles automáticos y una revisión por parte de un experto de la plataforma.
- **Revisión y retroalimentación:** el equipo de CRAN revisará el paquete y proporcionará retroalimentación útil para mejorarlo si es necesario.

Una vez el paquete cumple con los estándares de CRAN y es aprobado por el equipo de revisión, estará disponible para otros usuarios de R en el repositorio¹⁷. Este proceso puede llevar un tiempo aproximado de entre dos semanas a un mes, en función de la demananda y las revisiones que se requieran.

Tutorial del Paquete de R *smoothy*

El siguiente documento conforma la *vignette* (tutorial) incluida en el paquete de R *smoothy*, disponible online¹⁸, en el que se describen las diferentes funciones que contiene y el proceso detallado para utilizar el algoritmo “smooth”.

¹⁶ Licencia de software libre y código abierto: https://es.wikipedia.org/wiki/GNU_General_Public_License

¹⁷ El enlace para el paquete smoothy: <https://cran.r-project.org/web/packages/smoothy/index.html>

¹⁸ <https://cran.r-project.org/web/packages/smoothy/vignettes/introduction-to-smoothy.html>

Introduction to Smoothy

Dan Ouchi

2023-08-06

Overview

The smoothy package is an R package that provides a collection of functions to smooth the drug exposition using electronic health records. Also, the package helps to prepare the electronic drug records to model the drug exposition and perform longitudinal treatment pattern analysis. This vignette serves as a guide to understanding the main functionalities of the package and provides examples of how to use the functions.

Installation

You can install the `smoothy` package from CRAN using the following command:

```
install.packages('smoothy')
```

Load packages:

```
library(smoothy)
library(tidyr)
library(knitr)
```

Functions

The `smoothy` package includes the following main functions:

- `smooth_algorithm()`: This function performs the smooth algorithm as described in Ouchi et.al. [\[ref\]](#)
- `smooth_parse()`: This function transforms the data to obtain the daily treatment.
- `smooth_deparse()`: This function returns to the original format but collapsing all the drugs exposed on the same day.
- `smooth_diff()`: This function computes the differences between the original treatment and the smoothed treatment.

The smooth algorithm

The algorithm described in the published paper [\[ref\]](#) is based on moving averages commonly used in time series analysis. Its purpose is to facilitate modeling of drug exposure using electronic drug records. The algorithm calculates the average or most probable treatment exposure for each day of the follow-up period.

The package contains all the necessary functions to execute the algorithm, which can be divided into three main steps:

- Transform the raw data (parse) with the `smooth_parse()` function
- Running the smoothing algorithm `smooth_algorithm()` function
- Collapsing the treatment periods from the smoothing algorithm with `smooth_deparse()` function

Additionally, the package includes one function to validate and better assess the impact of the algorithm on the original records:

- `smooth_diff()`: Calculates the differences between the original data and the smoothed data

Input dataset

The raw data (input data) should be structured as a data frame with columns representing unique identifiers for each drug administration event (id), the start date of drug administration (start_date), the end date of drug administration (end_date), and the name of the drug administered (drug).

Example dataset

The *smoothy* package comes with an example dataset called *drugtreatment*. This dataset contains information of 387 patients and the prescribed drugs. It has the following variables:

- idp: Identifier for each individual.
- start_date: Date of drug initiation.
- end_date: Date of drug end.
- drug: Type of drug exposed, coded as A, B, or C.

You can load the example dataset using the following command:

```
data(drugtreatment)
```

id	start_date	end_date	drug
867e99f2-6cc1-492c-a1c0-e36f48c7aa95	1971-01-02	1971-01-04	C
867e99f2-6cc1-492c-a1c0-e36f48c7aa95	1971-01-05	1971-03-09	C
867e99f2-6cc1-492c-a1c0-e36f48c7aa95	1971-03-10	1971-04-07	C
867e99f2-6cc1-492c-a1c0-e36f48c7aa95	1971-04-08	1971-04-28	C
867e99f2-6cc1-492c-a1c0-e36f48c7aa95	1971-04-29	1971-06-05	C
867e99f2-6cc1-492c-a1c0-e36f48c7aa95	1971-06-06	1971-08-24	C

The workflow

For this example, we are going to use a single patient:

```
my_data <- dplyr::filter(drugtreatment, id == "62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956")
```

Step 1. Transform the Original Data (parse)

We will expand the original drug records to represent daily active treatment during the defined study period.

```
structured_df <- smooth_parse(  
  id = my_data$id,  
  start_date = my_data$start_date,  
  end_date = my_data$end_date,  
  drug = my_data$drug,  
  study_from = "1970-01-01",  
  study_to = "1975-01-01"  
)
```

The *smooth_parse()* function will return a new data frame with four columns:

id	day	N	treatment
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-01	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-02	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-03	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-04	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-05	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-06	0	None

- id: unique identifier
- day: date vector with the day of follow-up
- N: Number of drugs exposed to in that day
- treatment: Character vector with the treatment. None: no exposed to any drug of study.

Step 2. Apply the Algorithm

Next, we will apply the smooth algorithm to the parsed data.

```
smoothed <- smooth_algorithm(  
  id = structured_df$id,  
  treatment = structured_df$treatment,  
  day = structured_df$day,  
  N = structured_df$N,  
  width = 61  
)
```

This function performs the algorithm on each patient and it returns the same data frame but with a new column **smoothed_treatment**:

id	treatment	day	N	smoothed_treatment
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	None	1970-01-01	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	None	1970-01-02	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	None	1970-01-03	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	None	1970-01-04	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	None	1970-01-05	0	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	None	1970-01-06	0	None

Step 3. Untransform the Smoothed Data (deparse)

To perform analysis on the treatment pattern, is better to collapse the treatments according to the exposition periods (end date and start date).

```
deparsed.sm <- smooth_deparse(smoothed$id, smoothed$day, smoothed$smoothed_treatment)
```

The `smooth_parse()` function use the smoothed data frame as input and returns a new one very similar to the original dataset that we are more commonly using in analysis. It has four columns:

id	start_date	end_date	treatment
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1970-01-01	1972-03-01	None
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1972-03-02	1972-04-13	A
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1972-04-14	1972-10-31	A+C
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1972-11-01	1973-06-21	A+B+C
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1973-06-22	1973-12-21	A+B
62eb1ebd-1b49-4ba7-9af9-ec0c4ada0956	1973-12-22	1975-01-01	None

- id: unique identifier.
- start_date: date vector with the day of treatment initiation.
- end_date: date vector with the day of treatment end.
- treatment: Character vector with the treatment (smoothed or not). None: no exposed to any drug of study.

Also, the function can be used on the non-smoothed treatment pattern (column *treatment*) which is the same as the initial data but including the untreated periods.

```
deparsed <- smooth_deparse(smoothed$id, smoothed$day, smoothed$treatment)
```

Step 4. Validation

The validation includes assessing the differences before and after applying the smooth algorithm, measuring the number of days changed and the proportion of total days, and visualizing the longitudinal treatment pattern to see the impact of the algorithm in the overall pattern.

4.1. Count Differences Before and After the Smooth Algorithm

We will calculate the number of days changed and the proportion of total days in overall, exposition period and at each treatment.

```
smooth_diff(treatment = smoothed$treatment, smoothed_treatment = smoothed$smoothed_treatment)
```

type	days_changed	proportion_of_change
Global	54	0.0295567
Expousure period	53	0.0804249
drug change A	1	0.0333333
drug change A+B	22	0.1073171
drug change A+B+C	0	0.0000000
drug change A+C	0	0.0000000
drug change C	31	1.0000000

4.2 Visualization

Using the following code, in ggplot2, we can visualize the longitudinal treatment pattern of each individual before and after using the smooth algorithm.

```
require(ggplot2)
#> Loading required package: ggplot2
require(gridExtra)
#> Loading required package: gridExtra

tts = unique(deparsed$treatment)
tts = tts[tts!='None']
yorder = c('None',tts[order(nchar(tts), tts)])
p = ggplot(deparsed, aes(x = start_date, y = treatment)) + geom_segment(aes(x = start_date,
  xend = end_date, y = treatment, yend = treatment), size = 2, alpha = 0.85,
  col = 'grey20') + theme_bw() + scale_x_date(date_breaks = '6 months') +
  scale_y_discrete(limits = yorder) + theme(axis.text.x = element_text(angle = 60,
  hjust = 1), legend.position = "none") + ylab("") + xlab("") +
  ggtitle(paste0('Original treatment'))
#> Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
#> i Please use `linewidth` instead.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.

tts = unique(deparsed.sm$treatment)
tts = tts[tts!='None']
yorder = c('None',tts[order(nchar(tts), tts)])
p.sm = ggplot(deparsed.sm, aes(x = start_date, y = treatment)) + geom_segment(aes(x = start_date,
  xend = end_date, y = treatment, yend = treatment), size = 2, alpha = 0.85,
  col = 'grey20') + theme_bw() + scale_x_date(date_breaks = '6 months') +
  scale_y_discrete(limits = yorder) + theme(axis.text.x = element_text(angle = 60,
  hjust = 1), legend.position = "none") + ylab("") + xlab("") +
  ggtitle(paste0('Smoothed treatment'))

grid.arrange(p,p.sm,ncol=1)
```

Choosing the windows size

The default window size is set to 61 days. This value is based on various simulation studies and real-world analyses, which have shown that 61 days effectively captures treatment patterns without omitting critical treatment periods.

However, it is essential to note that this window size may not be suitable for all patients. For cases where there are significant changes in the treatment pattern, it may be beneficial to manually review and assess whether the algorithm oversimplifies the treatment pattern.

In the following example, we have selected a patient with frequent treatment changes to illustrate the effects of different window sizes:

```
my_data <- dplyr::filter(drugstreatment, id == '25094328-3819-4061-941d-191c4e0bc939')

structured_df <- smooth_parse(
  id = my_data$id,
  start_date = my_data$start_date,
  end_date = my_data$end_date,
  drug = my_data$drug,
  study_from = "1970-01-01",
  study_to = "1975-01-01"
)

# smooth
smoothed <- smooth_algorithm(id = structured_df$id,
                             treatment = structured_df$treatment,
                             day = structured_df$day,
                             N = structured_df$N,
                             width = 31)
smoothed45 <- smooth_algorithm(id = structured_df$id,
                               treatment = structured_df$treatment,
                               day = structured_df$day,
                               N = structured_df$N,
                               width = 45)
smoothed61 <- smooth_algorithm(id = structured_df$id,
                               treatment = structured_df$treatment,
                               day = structured_df$day,
                               N = structured_df$N,
                               width = 61)

deparsed <- smooth_deparse(smoothed$id, smoothed$day, smoothed$treatment)
deparsed.sm <- smooth_deparse(smoothed$id, smoothed$day, smoothed$smoothed_treatment)
deparsed.sm45 <- smooth_deparse(smoothed45$id, smoothed45$day, smoothed45$smoothed_treatment)
deparsed.sm61 <- smooth_deparse(smoothed61$id, smoothed61$day, smoothed61$smoothed_treatment)
```

Utilizing the Algorithm in a Large Dataset

Executing the entire workflow in a real study can be computationally intensive. Without a workstation equipped with ample RAM and high-speed processors, applying the smooth algorithm without parallel computing may prove to be unfeasible.

To overcome this challenge, we demonstrate how to run the algorithm on the entire example dataset by dividing it into manageable chunks and parallelizing the calculations in R.

To achieve this, make sure to have the following R packages installed:

```
library(doParallel)
library(foreach)
library(snow)
library(doSNOW)
```

Below is a script demonstrating how to run the entire workflow in parallel in R:

1) Split Data in Chunks

In this example, the chunk size is 50, but in a big data set we recommend to use chunks of size 1000.

```
data = drugstreatment

# starting date:
s0 = Sys.time()

# create chunks:
chunksize = 50
niter <- n_distinct(drugs$id)
chunks <- ceiling(niter/chunksize)
inds <- split(seq_len(niter), sort(rep_len(seq_len(chunks),
                                         niter)))
```

2) Sockets, Cores and Progress bar

Parameters to pass into the `foreach()` function. Additionally, the progress bar feature provides a helpful visual representation of the ongoing computation, allowing you to monitor the progress and estimate the remaining time until completion

```
# parallel - cores and socket:
n.cores <- 3
cl <- snow::makeSOCKcluster(n.cores)
doSNOW::registerDoSNOW(cl)

# progress bar:
pb <- utils::txtProgressBar(min = 1, max = chunks, style = 3)
progress <- function(n) utils::setTxtProgressBar(pb,n)
opts <- list(progress = progress)
```

3) Parallel Execution using foreach() and %dopar%

The entire workflow is encompassed within the foreach() function, executed in a loop across all the data chunks using %dopar%. During each iteration, the computed output is saved in the tempdir() location in the RDS format. This allows for parallel processing, optimizing the performance and enhancing computational efficiency.

```
# path to temporal directory:
tmp.path = tempdir()
diff=FALSE

l <- foreach(c = 1:chunks,
  .packages = c("Kendall", "smoother", "data.table", "anytime", "dplyr"),
  .options.snow = opts,
  .multicombine = F) %dopar% {

  chunk.id <- unique(data$id)[inds[[c]]]

  # run the workflow in each individual from the chunk:
  df <- filter(data, id %in% chunk.id)

  # 1) parse data:
  structured_df <- smooth_parse(
    id = df$id,
    start_date = df$start_date,
    end_date = df$end_date,
    drug = df$drug,
    study_from = "1970-01-01",
    study_to = "1975-01-01"
  )

  # 2) smooth algorithm:
  width <- 61

  smoothed <- smooth_algorithm(
    id = structured_df$id,
    treatment = structured_df$treatment,
    day = structured_df$day,
    N = structured_df$N,
    width = width
  )

  # 3) deparse data (original format):
  deparsed_smoothed <- smooth_depense(
    smoothed$id,
    smoothed$day,
    smoothed$smoothed_treatment
  )

  # 4) Per patient changes due to smooth algorithm:
  if(diff){

    # Calculate differences by patient mapping with the group_map function:
    df <- smoothed %>%
      group_by(id) %>%
      group_map(~ smooth_diff(.treatment, .smoothed_treatment)) %>%
      bind_rows(.id = "group_id") %>%
      data.frame

    # Format output and filter global, exposure period:
    df <- df %>%
      mutate(percentage_of_change = round(proportion_of_change*100,2)) %>%
      filter(type%in%c('Global','Exposure period')) %>%
      mutate(type = factor(type,levels=c('Global','Exposure period'),
        labels=c('total_change','exposure_change')))

    # add 'id' and reshape:
    df <- df %>%
      left_join(data.frame(id=unique(smoothed$id),group_id = as.character(seq(1,n_d
      reshape2::dcast(id~type,value.var='percentage_of_change'))

    # attach to deparsed_smoothed dataframe:
    deparsed_smoothed <- left_join(
      deparsed_smoothed,
      df
    )

    rm(df)
  }

  # Save chunk output to a temporary folder:
  saveRDS(deparsed_smoothed,paste0(tmp.path,"/chunk_",c,".rds"))

  rm(df,structured_df,smoothed,deparsed_smoothed);gc()
}
```

3.1) Patient-Level Differences

Within the loop, there is an optional code segment used to calculate the differences after applying the smooth algorithm at a patient level. This step is entirely optional, but it can be useful for validation, especially in cases where patients exhibit significant changes (approximately 10% of change).

```
# Calculate differences by patient mapping with the group_map function:
aux <- smoothed %>%
  group_by(id) %>%
  group_map(~ smooth_diff(.$treatment,.$smoothed_treatment)) %>%
  bind_rows(.id = "group_id") %>%
  data.frame

# Format output and filter global, exposure period:
aux <- aux %>%
  mutate(percentage_of_change = round(proportion_of_change*100,2)) %>%
  filter(type%in%c('Global','Exposure period')) %>%
  mutate(type = factor(type,levels=c('Global','Exposure period'),
                        labels=c('total_change','exposure_change')))
# add 'id' and reshape:
aux <- aux %>%
  left_join(data.frame(id=unique(smoothed$id),group_id = as.character(seq(1,n_distinct(smoothed$id))))
  reshape2::dcast(id~type,value.var='percentage_of_change')

# join to deparsed_smoothed dataframe:
deparsed_smoothed <- left_join(
  deparsed_smoothed,
  aux
)
```

After completing the process, it is essential to close the clusters to release the computing resources. Additionally, you may want to calculate the total computation time for the entire workflow.

```
# close sockets:
close(pb)
snow::stopCluster(cl)

# Time to finish the process:
t0 = Sys.time() - s0
cat("\n The process finished in", round(t0), units(t0))
```

4) Read and Combine All Chunks

To complete the process, we read the .rds files stored in the temporary folder and combine them into a single data.frame using the bind_rows function:

```
# Import and combine all chunks into a single data.frame:
rds_files <- list.files(tmp.path, pattern = "chunk_", full.names = TRUE)
all_chunks <- bind_rows(lapply(rds_files, readRDS))
```

The computation time may vary depending on the number of samples and the study period. We conducted several tests on both Mac and Windows platforms, and the performance was similar.

As a reference:

- PC/Mac with 8 cores and 16GB of RAM
- A dataset with 500k patients and 2.8M electronic prescriptions
- Follow-up period of 6 years

The entire process took 16 hours to finish.

